EKSAKTA
Journal of Sciences and Data Analysis

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI: 10.20885/EKSAKTA.vol1.iss1.art11

Research Article

A Comparison between Nonparametric Approach: Smoothing Spline and B-Spline to Analyze The Total of Train Passengers in Sumatra Island

Drajat Indra Purnama[1,*]

[1] Badan Pusat Statistik (BPS) Kabupaten Parigi Moutong, Provinsi Sulawesi Tengah, Indonesia
[*] Corresponding author: drajatindrapurnama@bps.go.id

Received: 7 January 2020; Accepted: 17 February 2020; Published: 20 February 2020

Abstract: The train is one of the means of land transportation of the most desirable communities other than ground transportation such as bus or car. This is because the rail has an advantage that is free from congestion. Increasing mobility of people, means of transportation that is free from congestion increasingly in demand. In Indonesia, the only railway in Java and Sumatra. Either in Java or Sumatra Island train passenger numbers have increased every year. Based on data from BPS-Statistics, the number of train passengers in Sumatra during the last five years has increased an average of 14.8 percent per year. Noparametric regression model that can be used to describe the pattern of data on the number of train passengers in Sumatra Island is smoothing spline regression and B-spline regression. The purpose of this study is to find the most nonparametric regression model to describe the pattern of the relationship between the time and number of train passengers on Sumatra Island. Smoothing spline and B-spline models were compared by looking at the regression curve and the value of *Mean Square Error* (MSE). The results of this study indicate that smoothing spline model is more appropriate to see the pattern of the relationship between the time and number of train passengers in Sumatra Island. This can be seen from the MSE of smoothing spline models 2,742.801 smaller than the MSE of B-spline models 3,847.657.

Keywords: spline, smoothing spline, B-spline, train in Sumatra island

Introduction

Train is a railroad-based land transportation tool that can transport people or goods on a larger scale compared to other land transportation facilities. In addition, the train has various advantages including being free from traffic because it has its own track, and is more fuel efficient. Railroad has an important role in population mobility. From the graph of the number of passengers and goods transported by railroad, public interest in rail transport services is seen increasing. Based on BPS-Statistics, data on the number of train passengers in Indonesia during the 2013-2018 period experienced an increase in the number of passengers an average of 14.6 percent annually.

PT Kereta Api Indonesia (PT. KAI) is the sole organizer of railroad services in Indonesia. PT. KAI is currently only available on Java and Sumatra. In Java, the operational area of PT. KAI consists of the Operations Area (DAOP) which is located starting from Daop I (Jakarta) to Daop IX (Jember). As different as in Java where the train connects almost all cities on Java, on the island of Sumatra the train only passes through a few regions. This is because the railroad tracks on the island of Sumatra are intermittent, aka not connecting major cities on the island of Sumatra. The area of railroad operations on Sumatra Island consists of four Regional Divisions (Divre) namely Divre I (North Sumatra and Aceh), Divre (West Sumatra), Divre III (Palembang) and Divre IV (Tanjungkarang). Nevertheless, the train on the island of Sumatra has an important role as a means of public transportation. This is proven that, during the period of 2013-2018, the number of passengers increased by an average of 14.8 percent annually. Therefore, it is necessary to analyze the increase or decrease in the number of train passengers on the island of Sumatra.

One method for analyzing data is regression analysis. In a regression analysis, if the pattern of the relationship between the predictor variable and the response variable is unknown form of the function, then a nonparametric regression approach is used. One nonparametric approach that is popular in is Spline [1].

Splines regression is one of the regression models built from segments in polynomial functions. Splines estimators have high flexibility compared to other estimators. Splines can also overcome data patterns with the help of knot points and produce smooth curves [2]. Research on spline has been widely developed, among others, carried out by Doksum and Koo [3], Wand [4] and Huang [5].

Spline regression commonly used for analysis is smoothing spline and B-spline. Smoothing spline can provide a better analysis result because it can overcome data patterns that show a sharp rise and fall and the resulting curve is relatively smooth [6]. The disadvantage of smoothing spline is that when the order of the spline is high and the greater the knot vector will form a matrix in the singular calculation, so that there may be a spline equation that cannot be solved. Problems with high order splines can be overcome with B-spline. But the difficulty with B-spline is that the B-spline basis can only be defined recursively and cannot be evaluated directly [7].

This research will compare the spline nonparametric regression model, namely the smoothing spline and B-spline approaches in the data analysis of the number of train passengers on the island of Sumatra from January 2006 to December 2018.

Materials and Methods

Nonparametric regression

Nonparametric regression is one of the statistical methods used to determine the pattern of relationships between predictor variables and response variables of unknown function. Wood [8] defines the relationship between predictor variables (X) and response variables (Y) in nonparametric regression modeling written

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1,2,\dots,n$$

where $f(x_i)$is a function of the unknown independent variable X and ε_i is the error of every i-th observation.

Nonparametric regression is used if the relationship pattern of the response variable and the predictor variable is unknown in the shape of the regression curve. In other words the form $f(x_i)$ and the number of parameters to be estimated in the nonparametric model are unknown at the beginning. In nonparametric regression the regression curve is assumed to be smooth. Estimation of nonparametric functions is based on observational data using *smoothing* techniques.

Smoothing splines regression

Smoothing splines regression is a matching curve with smoothing for a set of observations using the splines function [8]. Estimation of $f(x_i)$ using *Ordinary Least Square* (OLS) produces a coarse function so that a finer function is needed so that the number of residual squares becomes small. The giving of a smoothing or penalty is used in smoothing splines regression in order to get a smooth function. The estimated splines function $f(x_i)$ is to minimize the *Penalized Least Square* (PLS) function written as

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

where$\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2$ is the sum of the squares of error, $\int\left[f''(x)\right]^2 dx$ is the *roughness penalty*, which is a measure of the smoothness of the curve, and λ is a smoothing parameter, the value of λ is between 0 and 1. If the value of λ approaches 1, it will give a large penalty weight (smoothness) and has a small variance.

B-spline regression

Spline is a segmented polynomial model (*piecewise polynomial*), which is a piece of polynomial that has segmented properties at intervals k formed at knot points [9]. Knot points are points that indicate changes in data in sub-intervals. The weakness of spline is when the spline order is high, many knots or knots are too close will form a matrix which is almost singular in calculation, so that normal equations cannot be solved. Another base that can overcome spline weakness is B-spline. The difficulty with B-splines is that the basis on B-splines can only be defined recursively so that they cannot be evaluated directly.

According to Eubank [7], B-splines are *piecewise polynomial* functions with local *support* for certain polynomial degrees. Jth B-splines with degrees v based on knots with u point knots, an additional knot is defined as much as $2v$ so a row of knots is obtained $T(t_1,\dots,t_v,t_{v+1},\dots,t_{u+v},t_{u+v+1},\dots,t_{u+2v})$ where $t_1 =$

$\cdots = t_v < t_{v+1} < \cdots < t_{u+v} < t_{u+v+1} = \cdots = t_{u+2v}$. So the jth B-spline for $j = 1, \ldots, v+u$ is denoted by a recursive formula written as

$$B_j(x;v) = \frac{x-t_j}{t_{j+v-1}-t_j} B_j(x;v-1) + \left(1 - \frac{x-t_{j+1}}{t_{j+v}-t_{j+1}} B_{j+1}(x;v-1)\right)$$

where

$$B_j(x;v) = \begin{cases} 1, \text{if} t_j \le x \le t_{j+1} \\ \\ 0, \text{the other} \end{cases}$$

Normalized B-spline means that $\forall x: \sum_{j=1}^{v+u} B_j(x;v) = 1$.

In regression of the mean $y_i = f(x_i) + \varepsilon_i,\ i = 1, \ldots, n$, always assumed that $f(x_i)$ is a smooth function that can be approximated by linear combinations of B-splines bases

$$f(x) \approx \sum_{j=1}^{m} \alpha_j B_j(x;v)$$

where$\alpha = (\alpha_1, \ldots, \alpha_j)$ is a coefficient vector of the base B-splines $B_j(.;v)$ with degrees of freedom v and $u+1$*equidistant* knots for $j = 1, \ldots, v+u = m$. The objective function of B-splines regression is written as

$$\hat{\alpha} = \text{argmin}_\alpha \left\{ \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{m} \alpha_j B_j(x_i;v) \right) \right\},$$

So the regression model becomes

$$y_i = \sum_{j=1}^{m} \alpha_j B_j(x_i;v), \quad i = 1,2, \ldots, n$$

Selecting lambda (λ) and optimal knot points

The formation of a nonparametric smoothing spline regression model must pay attention to the value of the smoothing parameter or lamda (λ). The smoothing splines estimator is very dependent on the smoothing parameters or lamda (λ)so the selection of smoothing parameters is important in finding the most appropriate smoothing splines estimator [7].

The best spline regression model depends on the optimal knot point [7]. Criteria that must be considered in forming a B-spline regression model are determining the order or *degree* for the regression model and the number of knots. Knots are joint fusion points where there is a change in behavior in the data. Too many knots tend to *overfit* and *rigid* B-spline curves, while knots that are too small make the curve *oversmooth* and are less able to describe the pattern of data distribution.

In Eubank [7], it is stated that the performance measure of the estimator of the regression function can be determined including the *Mean Square Error* (MSE) and *Generalized Cross Validation* (GCV). GCV was introduced by Craven and Wahba[10] in the context of smoothing splines. Criteria for selecting the minimum GCV value parameters according to Lee [11] are obtained by the formula written as

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

where

I : identity matrix
$A(k)$: matrix$X(X^TVX)^{-1}X^TV$
$MSE(GCV)$: $\frac{\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2}{n}$

Methods

Data Sources and Research Variables

The data used in this study are secondary data from BPS-Statistics [12] for the period January 2006 - December 2018. These data are monthly data with time as a predictor variable (X) and the number of train passengers on the island of Sumatra as a response variable (Y). Data processing and analysis in this study using the help of *R* series 3.5.1 software.

Research Stages

The steps of data analysis in this study are

1. Conduct descriptive analysis on data
2. Make a *scatter plot* between the predictor variable and the response variable
3. Smooting spline regression approach
   a. Calculating the value of lamda(λ)
   b. Determine the optimal value of lamda(λ) based on GCV criteria
   c. Make a smoothing spline regression curve
   d. Calculates MSE smoothing spline regression
4. B-spline regression approach
   a. Calculate knots value
   b. Determine optimal knot values based on *Adj* GCV criteria
   c. Make a B-spline regression curve
   d. Calculates MSE B-spline regression
5. Comparing curves and MSE smoothing spline regression and B-spline regression.
6. Conclusion

Results and Discussion

Descriptive analysis

The data processed and analyzed in this study are monthly data from BPS- Statistics, in the form of the number of train passengers on the island of Sumatra between January 2006 and December 2018. This means that this study uses the number of observations over 156 months. The following is descriptive of the data on the number of train passengers on the island of Sumatra, which is presented in Table 1.

Table 1. Descriptive Data

Description	Passenger AmountKA on Sumatra Island (Thousand People)
Minimum	210
Median	395
Mean	414.2
Maximum	768
Standard Deviation	120.355

From Table 1. the information is obtained that the number of train passengers on the island of Sumatra has the highest value of 768 thousand people (December 2018), the lowest value of 210 thousand people (February 2007) and an average passenger of 414.2 thousand people each month.

Scatter plot between time and number of train passengers on the island of Sumatra

The pattern of the relationship between time as a predictor variable to the number of train passengers on the island of Sumatra as a response variable is shown in Figure 1. Based on Figure 1 it can be seen that the relationship of time and number of train passengers on the island of Sumatra is spread with a down and up pattern, but it is still difficult to determine the exact relationship between the two variables. This indicates that there is a nonparametric component. Therefore it can be continued in the nonparametric smoothing spline and B-spline regression models.

E-ISSN: 2720-9326
P-ISSN: 2716-0459

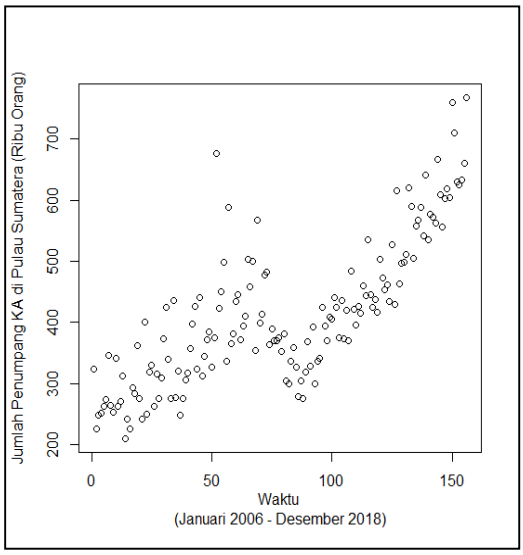

Figure 1. Scatter plot relationship between time and number of train passengers on Sumatra island

Smooting spline regression

Nonparametric smoothing spline regression is one of the methods used to estimate nonparametric regression curves. The criteria that must be considered in establishing a nonparametric smoothing spline regression model are the smoothing parameter or lamda (λ) which is an important part in determining the smoothing splines regression estimator. Lamda (λ) is used as a control tool to determine the smoothness of a curve. Determination of optimal lambda based on GCV values.

Table 2. presents the value of the smoothener or lamda (λ) that was tested on the smoothing spline regression. It can be seen that the optimal parameters or lamda (λ) are with a GCV value of 3,169.with lamda (λ)of 2,608 x 10^{-4}.

Table 2. Lambda Values (λ) and GCV

Lambda (λ)	GCV
5 x 10^{-1}	5,226.939
5 x 10^{-2}	4,140.463
2.608 x 10^{-4}	3,169.087
5 x 10^{-3}	3,434.372
5 x 10^{-4}	3,179.703
5 x 10^{-5}	3,222.322
5 x 10^{-6}	3,437.019
5 x 10^{-7}	3,952.639
5 x 10^{-8}	5,289.328

Then a smoothing splines regression curve is formed with an optimal smoothing parameter or lambda (λ) of 2.608 x 10^{-4} which can be seen in Figure 2. The smoothing spline curve is influenced by the value of lamda (λ), so the resulting curve is getting coarser. The greater the value of lamda (λ), the curve will be smooth, but the results obtained are not necessarily good.

In Figure 2., it can be seen that the smoothing spline regression curve can be said to be good because the regression curve is smooth and approaches the shape of the plot with a small bias.

E-ISSN: 2720-9326
P-ISSN: 2716-0459

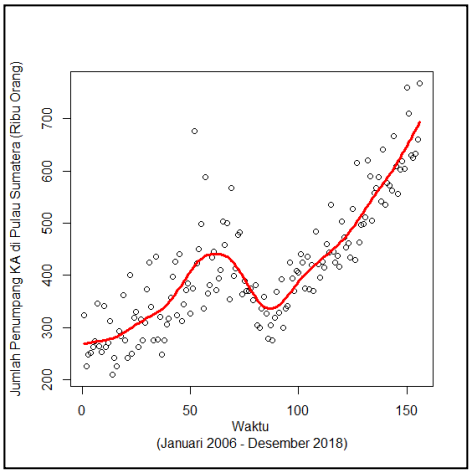

Figure 2. Smoothing spline regression curve with lamda(λ)= 2,608 x 10^{-4}.

B-spline regression

B-spline regression is used to resolve the weakness of the spline model in high order. Criteria that must be considered in forming a B-spline regression model are determining the order or *degree* for the regression model and the number of knots. Knots are joint fusion points where patterns of behavior change at different intervals. In this study the order used only at degree 3. Whereas for the selection of optimum number of knots the *Adj* GCV value was used.

Table 3. Number of Knots, Optimal Knot Values and *Adj* GCV B-Spline

Number of Knots	Optimal Knot					*Adj* GCV
1	36.38139					4,068.82782
2	72.99968	74.33118				3,540.90107
3	83.99188	85.99998	87.03440			3,502.79177
4	42.07558	74.33543	86.99794	88.55600		3,612.67783
5	42.20109	74.12072	86.99710	88.01065	153.98800	3,792.56936

The minimum *Adj* GCV value from the B-spline regression model with *degree* 3 and several knots points is presented in Table 3. Based on table 3, the minimum *Adj* GCV value is obtained at 3 knots as many as 3,502.79177 and the optimal knots at 83.99188, 85.99998 and 87.03440. Using *degree* 3 and number of knots 3 with optimal knots of 83.99188, 85.99998 and 87.03440, a B-spline regression curve can be determined which can be seen in Figure 3.

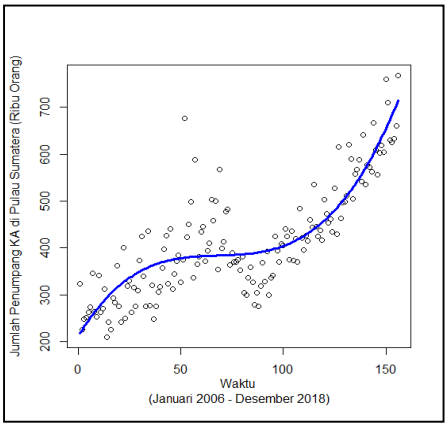

Figure 3. B-spline regression curve with degree 3 and optimal knots

E-ISSN: 2720-9326
P-ISSN: 2716-0459

In B-spline regression, the more number of knots, the B-spline *overfit* curve. From Figure 3 it can be seen that the B-Spline *degree* 3 regression with optimal knots of 83.99188, 85.99998 and 87.03440 can model the number of train passengers on Sumatra Island quite well because it has a smooth regression curve and is able to reach the data distribution.

Comparing Smoothing Spline Regression and B-Spline Regression

A comparison between smoothing spline regression curves and B-spline regression curves is presented in Figure 4. It can be seen that both the smooting spline curve and the B-splines curve almost cover all available data distribution. However, the B-spline curve is smoother than the smoothing spline curve. Although the B-spline curve is coarser, it can be seen that the B-spline curve is more able to show fluctuating up and down data during the study period.

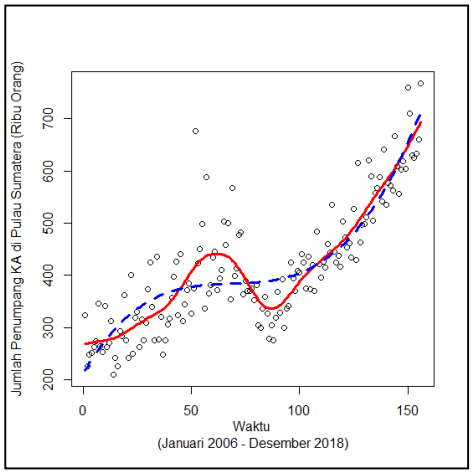

Figure 4. Comparison between smoothing spline regression curve and B-spline regression curve

The comparison of MSE values between the smoothing spline and B-spline models is presented in Table 4. It is seen that the MSE value of the smoothing spline model is smaller than the MSE B-spline model. So, in this study it can be stated that the pattern of the relationship between time and number of train passengers on the island of Sumatra is more appropriate using the nonparametric smoothing spline model.

Tabel 4. Comparison of Smoothing Spline and B-Spline

Model	MSE
Smoothing Spline	2.742,801
B-SPline	3.847,657

Conclusion

In this study, the smooting spline curve model and the B-splines curve almost cover all available data distribution. B-spline curves are smoother than smoothing spline curves, but B-spline curves are more able to show fluctuations in rising and falling data during the study period. The exact nonparametric regression model used to describe the relationship between time and number of train passengers on the island of Sumatra in this study is the smoothing spline model. It is based that the smoothing spline model has a smaller MSE value.

References

[1] G. Wahba, Spline Models for Observational Data, SIAM, CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia, 1990, Vol. 59.
[2] W. Hardle, Applied Nonparametric Regression, Springer-Verlag, Berlin, 1994.
[3] K. Doksum, Y.J. Koo, On Spline Estimators and Prediction Intervals in Nonparametric Regression, Computational Statistics and Data Analysis 35 (2000) 67 – 82.

[4] M. P. Wand, A Comparison of Regression Spline Smoothing Procedures, Departments of Biostatistics, School of Public Health, Harvard, 2005.
[5] Z. J. Huang, Local Asymptotic for Polynomial Spline Regression, The Annual Statistics 31 (2003) 1600-1635.
[6] T. J. Hastie, R.J. Tibshirani, Generalized Additive Models, Chapman & Hall, London, 1990.
[7] R. Eubank, Spline Smoothing and Nonparametric Regression, Marcel Dekker, New York, 1988.
[8] N. S. Wood, Generalized Additive Models: an Introduction With R, Boca Raton: Chapman & Hall/CRC, 2006.
[9] J. Wang, L, Yang, Polynomial Spline Confidence Bands for Regression Curves, Statistica Sinica 19 (2009) 325-342.
[10] P. Craven, G. Wahba, Smoothing Noisy Data with Spline Functions: Estimating the Correct, Degree of Smoothing by the Method of Generalized Cross-Validation, Numer Math University of Wisconsin. 31 (1979) 377-403.
[11] T. C. M. Lee, Smoothing Parameter Selection for Smoothing Splines: a Simulation Study, Computational Statistic & Data Analysis 42 (2003) 139-148.
[12] BPS-Statistics. https://www.bps.go.id/linkTableDinamis/view/id/815, accessed on May 19, 2019.