MODELING AIR QUALITY INDEX IN INDONESIA USING SMOOTHING SPLINES AND TRUNCATED SPLINES REGRESSION

Nadhia Az Zahra[1)]
Khoirin Nisa[2)]
Misgiyati[3)]
Nusyirwan[4)]
[1,2,3,4)] Mathematics Study Program, Faculty of Mathematics and Natural Sciences, University of Lampung, Bandar Lampung, Indonesia
e-mail: khoirin.nisa@unila.ac.id

ABSTRACT

The Air Quality Index (AQI) is a composite indicator that reflects regional air quality conditions and is influenced by multiple determinants with complex and nonlinear relationships. In such circumstances, parametric regression may be restrictive because it requires a predetermined functional form. This study applies spline based nonparametric regression using smoothing splines and truncated splines to model AQI in Indonesia and to compare the performance of both approaches. AQI is treated as the response variable, while population density, land cover area within and outside forest areas, and the number of motor vehicles are considered as predictor variables. For smoothing splines, the optimal smoothing parameter is selected using Generalized Cross Validation, whereas truncated splines are estimated using Ordinary Least Squares under various knot configurations and selected based on the minimum Generalized Cross Validation value. Model performance is evaluated using Generalized Cross Validation, Mean Squared Error, and Adjusted R squared. The study aims to identify the most appropriate model and to determine key factors influencing AQI variation in Indonesia, thereby providing empirical support for environmental policy making. The results show that the smoothing spline model provides better performance than the truncated spline model, with a lower Mean Squared Error (MSE) of 0.0716 and a higher Adjusted R^2 of 0.794. These results indicate that smoothing splines are more effective in capturing the nonlinear relationships influencing AQI variation in Indonesia.

Keywords: air quality index (AQI), nonparametric regression, smoothing splines, truncated splines, generalized cross validation (GCV).

INTRODUCTION

Air Quality Index (AQI) data represent an important environmental indicator used to describe the level of air pollution in a certain area. AQI is constructed from several pollutant parameters, such as NO_2, SO_2, PM_{10}, $PM_{2.5}$, CO, and O_3, which are combined into a single index value that is easily understood by the public and policy makers. This index is commonly classified into several categories ranging from good to hazardous, allowing communities to quickly recognize air quality conditions and take appropriate health precautions. In Indonesia, AQI data are crucial for monitoring environmental quality, supporting public health protection, guiding urban planning, and evaluating the effectiveness of environmental policies. Accurate modeling of AQI is therefore essential to understand its pattern, identify influencing factors, and provide scientific input for air quality management.

Regression analysis is a statistical method used to examine the relationship between a response variable and one or more predictor variables. In practice, regression analysis can be

carried out using parametric, nonparametric, and semiparametric approaches (Dani & Adrianingsih, 2021). Parametric regression is appropriate when the relationship between variables follows a known functional form, whereas nonparametric regression is more suitable when the data pattern is unclear or complex (Wongkar et al., 2023). In such situations, nonparametric regression is often preferred because of its flexibility in adapting to various data patterns.

One of the well-known approaches in nonparametric regression is spline regression. Splines are effective because they can accommodate changes in data behavior across different sub-intervals and are widely used because of their flexibility, smoothness, and ease of interpretation (Handayani et al., 2024; Mariati et al., 2021). Several spline estimation methods exist, including truncated splines and smoothing splines (Mariati et al., 2021). Among these, smoothing splines are useful for modeling smooth data patterns by balancing goodness of fit and curve smoothness through a smoothing parameter, while truncated splines are advantageous for capturing local characteristics of data through piecewise polynomial functions defined over segmented intervals (Ariesta et al., 2021; Suparti et al., 2018; Xu & Wang, 2021).

Several previous studies have applied nonparametric regression using smoothing and truncated splines. Mariati et al. (2021) used smoothing splines in social data analysis, demonstrating their ability to model complex trends. Nurcahayani et al. (2019) applied truncated splines to model average years of schooling in districts across Java, showing their effectiveness in capturing local patterns. Comparative studies, such as Fatmawati et al. (2019), reported that smoothing splines outperformed truncated splines in modeling human blood pressure data based on Mean Squared Error (MSE). However, studies comparing smoothing splines and truncated splines for modeling the Air Quality Index in Indonesia are still limited, particularly those that evaluate their relative performance using provincial-level environmental and demographic predictors. Therefore, further empirical evidence is needed to determine which spline approach is more appropriate for modeling AQI variation in Indonesia.

Nonparametric spline regression, including smoothing and truncated splines, is particularly suitable for analyzing social and environmental data, where relationships among variables are often complex and nonlinear. AQI is influenced by various factors, including population density, land cover, and the number of vehicles. High population density is usually associated with intense human activities, leading to increased emissions (Pant et al., 2023). Land cover influences the environment's ability to absorb and disperse pollutants, and its effect may not be proportional because differences in vegetation extent and distribution can lead to different levels of pollutant reduction (Sahani et al., 2024). In addition, vehicles are a major source of air pollution, especially in urban areas with heavy traffic, and their effect on AQI may also be nonlinear because emissions can increase more rapidly under congested traffic conditions. Thus, the relationship between these factors and AQI is often not linear, and a flexible modeling approach is needed to capture such patterns appropriately.

Based on this background, this study applies smoothing spline and truncated spline models to estimate the Air Quality Index in Indonesia and compares the performance of both methods in modeling AQI. This study contributes by providing a direct comparison of the two spline-based nonparametric approaches using Generalized Cross Validation (GCV), Mean Squared Error (MSE), and Adjusted R^2, as well as by identifying the model that is more suitable for explaining AQI variation in Indonesia.

METHOD

Type of Research

This study used secondary data obtained from Statistics Indonesia for 2023, consisting of 32 observations representing provinces in Indonesia. The response variable was the Air Quality

Index (AQI), while the predictor variables were population density, land cover area, and the number of motor vehicles. Prior to model estimation, all variables were standardized to account for differences in measurement scales. The analysis was conducted using R statistical software with spline-related packages.

Data Analysis Technique

The data analysis in this study follows these steps:

1) Descriptive analysis of the AQI and predictor variables (population density, land cover, and motor vehicle numbers).
2) Exploratory Data Analysis (EDA) to assess relationships between the variables using visualizations and correlations.
3) Nonparametric regression modelling using:
   a. Smoothing Splines: Optimal smoothing parameter selected via Generalized Cross Validation.
   b. Truncated Splines: Best knot configuration selected based on Generalized Cross Validation.
4) Model evaluation using Generalized Cross Validation, Mean Squared Error, and Adjusted R^2.
5) Model comparison to determine the best-fitting model.
6) Interpretation of results to identify factors influencing AQI.

Nonparametric Regression

With the advent of the latter part of the 1990s, the use of nonparametric regression as a data analysis approach witnessed a significant growth. This development finds impetus from the requirements for methods to effectively handle heterogeneous data that often fail to meet the parametric assumptions. As a highly dynamic approach, this methodology assumes great importance in the study of longitudinal data (Sriliana et al., 2022).

Common situations for applying nonparametric regression occur when the form of the regression curve is not predetermined or when there is minimal prior insight into the structure of the data pattern (Chen et al., 2021).

This type of regression is particularly useful when the shape of the data is unknown or when there is limited prior knowledge regarding the data pattern (Sriliana et al., 2022). A widely used method within nonparametric regression is splines regression, which models data flexibly by fitting piecewise polynomial functions.

Smoothing Splines

Smoothing spline regression is a nonparametric approach that estimates a smooth function to represent the relationship between variables. This method fits the data while controlling the level of smoothness so that the resulting model does not become overly complex (Centofanti et al., 2023).

The smoothing splines estimator $\hat{f}_\lambda$ for *f* is defined as the one that minimizes the following Penalized Least Square (PLS) function (Nisa et al., 2018).

$$\hat{f}_\lambda = arg\min_f \left\{ \sum_{i=1}^{n} \left[y_i - f(x_i) \right]^2 + \lambda \int_0^1 \left[f^{"}(x) \right]^2 dx \right\} \quad (1)$$

where $\lambda \geq 0$is the smoothing parameter that controls the trade-off between goodness of fit and smoothness. The first term measures the fit of the model to the data, while the second term is a roughness penalty that controls the curvature of the estimated function.

Truncated Splines

Truncated spline regression is a nonparametric method that models relationships using piecewise polynomial functions connected at knot points. The knots allow the regression curve to capture local changes in the data pattern. A truncated spline function of order m with knot points $K_1, K_2, \dots, K_r$ can be written as:

$$f(x_i) = \beta_0 + \sum_{j=1}^{m} \beta_j x_i^j + \sum_{l=1}^{r} \gamma_l (x_i - K_l)_+^m, \quad (2)$$

where

$$(x_i - K_l)_+^m = \begin{cases} (x_i - K_l)^m, x_i \geq K_l \\ 0 \quad\quad\quad , x_i < K_l. \end{cases} \quad (3)$$

In this representation, $\beta_j$are the polynomial coefficients, $\gamma_l$are the coefficients associated with the truncated basis functions, and $K_l$denotes the knot locations.

Applications of Regression Models

Regression models have been widely applied in various fields, including environmental science, economics, and social sciences. For example, in environmental studies, regression models are used to predict air quality, as seen in studies focusing on the Air Quality Index (AQI), which is influenced by various factors such as population density, land cover, and the number of vehicles (Pant et al., 2023).

Smoothing splines and truncated splines have proven effective in such analyses. Smoothing splines provide a smooth fit to the data while balancing the trade-off between smoothness and model fit (Xu & Wang, 2021). Truncated splines, on the other hand, divide the data into segments and fit a polynomial within each segment. This approach is useful for modeling data with abrupt changes or local variations (Suparti et al., 2018).

Mariati et al. (2021) used smoothing splines for modeling social data trends, while Nurcahayani et al. (2019) applied truncated splines to model education data, demonstrating the flexibility of these nonparametric methods for handling complex datasets.

Model Evaluation

Model performance was evaluated using Generalized Cross Validation (GCV), Mean Squared Error (MSE), and Adjusted R^2. GCV was used to select the optimal smoothing parameter and knot configuration, while MSE and Adjusted R^2 were used to compare the predictive accuracy and explanatory power of the final models.

RESULTS AND DISCUSSION

Characteristics of Air Quality Index Data

The descriptive statistics of AQI and the predictor variables used in this study are presented in Table 1.

Table 1. Descriptive Statistics

Variable	Mean	Standard Deviation	Minimum	Q1	Q2	Q3	Maximum
Y	90.916	3.042	81.39	90.01	90.92	92.83	96.22
X1	254	366.93	10	48	101	224	1346
X2	5002.384	3779.43	321.9	2010.2	4296.8	6306.8	15192.9
X3	162006.9	190593.25	9696	51241	81303	198811	782173

Table 1 shows that AQI has a mean of 90.916 with relatively low variability across provinces. In contrast, the predictor variables exhibit substantial variation, particularly population density and the number of motor vehicles. Because the variables were measured on different scales, all variables were standardized prior to model estimation.

Table 2. Descriptive Statistics of Standardized Variables

Variable	Mean	Standard Deviation	Minimum	Q1	Q2	Q3	Maximum
Y	0	1	-3.131	-0.298	0.001	0.630	1.743
X1	0	1	-0.664	-0.561	-0.417	-0.081	2.975
X2	0	1	-1.238	-0.791	-0.186	0.345	2.696
X3	0	1	-0.799	-0.581	-0.423	0.193	3.253

As shown in Table 2, the standardized variables have a mean of 0 and a standard deviation of 1, confirming that the scaling procedure was applied consistently. The standardized values facilitate comparison across variables and improve model estimation by reducing the effect of differences in measurement units.

Figure 1 presents the distributions of the standardized predictor variables. Population density and the number of motor vehicles show right-skewed distributions, indicating that several provinces have substantially higher values than the others. Land cover appears more evenly distributed, although some variation remains across provinces. These differences suggest that the relationships between AQI and the predictor variables may not be adequately represented by a simple linear model.

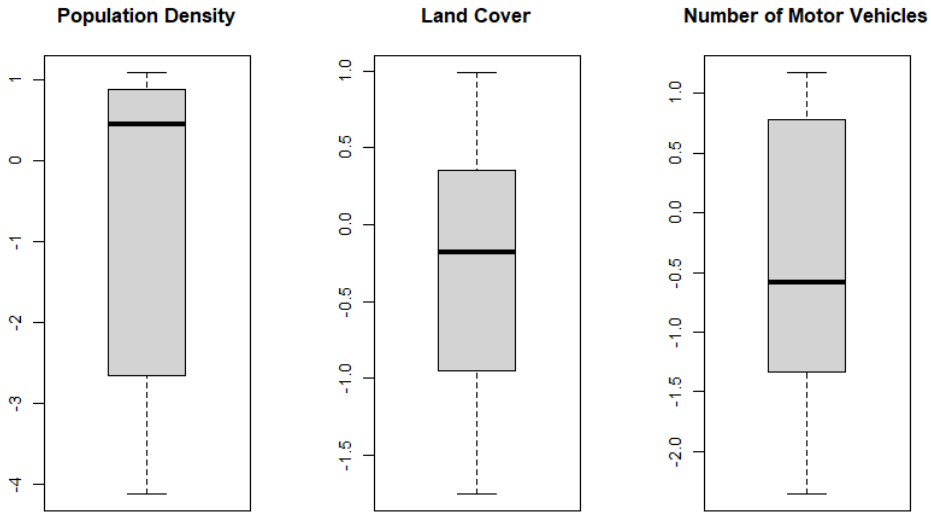

Figure 1. Boxplot of Population Density, Land Cover, and Number of Motor Vehicles

Scatterplot of Air Quality Index

To examine the correlation between the Air Quality Index (AQI) and predictor variables, scatterplots were used to visualize the distribution patterns and relationships, showing if they are increasing, decreasing, or have no clear trend.

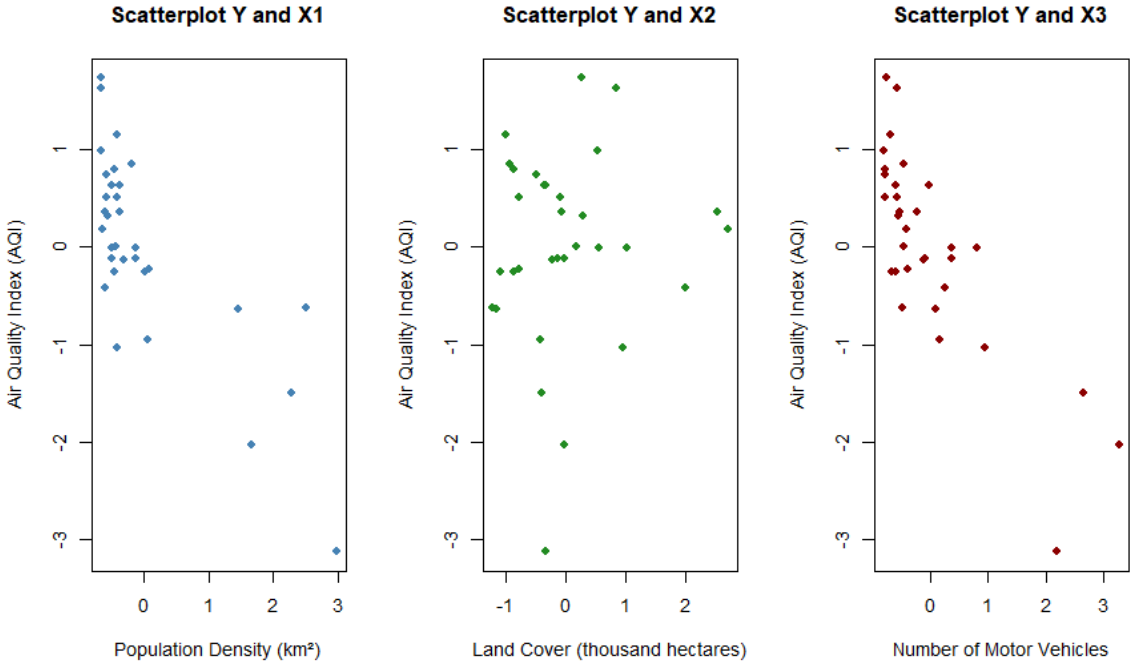

Figure 2. Scatterplots Showing the Relationship Between AQI and Predictor Variables

Based on Figure 2, the scatterplot illustrates the relationship between the Air Quality Index (AQI) and population density, where the data points are widely dispersed and do not form a clear linear pattern. Most provinces are characterized by relatively low population density with higher

AQI values, whereas provinces with very high population density tend to experience a noticeable decline in AQI, indicating a nonlinear relationship.

Similarly, the scatterplots for land cover and the number of motor vehicles also show no clear linear pattern with AQI. The observed dispersion and curvature suggest that a flexible nonparametric approach is more appropriate than a simple linear regression model for capturing the underlying relationships.

Table 1. Evaluation of the Simple Linear Regression Model

Statistical Measure	Value
SSE	7.948
SST	31.0
R^2	0.716
MSE	0.248

The linear model was not sufficient to capture the observed patterns in the data, particularly because the scatterplots suggested nonlinear relationships between AQI and the predictor variables. Therefore, spline-based nonparametric regression was employed to provide a more flexible representation of these relationships.

Smoothing Splines

In smoothing spline regression, the smoothing parameter (λ) controls the trade-off between model fit and smoothness. A smaller λproduces a more flexible curve, whereas a larger λ results in a smoother curve. The optimal value of $\lambda$was selected by minimizing the Generalized Cross Validation (GCV) criterion.

1) Determination of the Optimal Smoothing Parameters

The smoothing splines method balances the model's fit and smoothness using the smoothing parameter (λ), where a smaller λ fits the data closely and a larger λ smooths the curve. The optimal λ is determined by minimizing the Generalized Cross Validation (GCV), and the corresponding values are shown in the figure below.

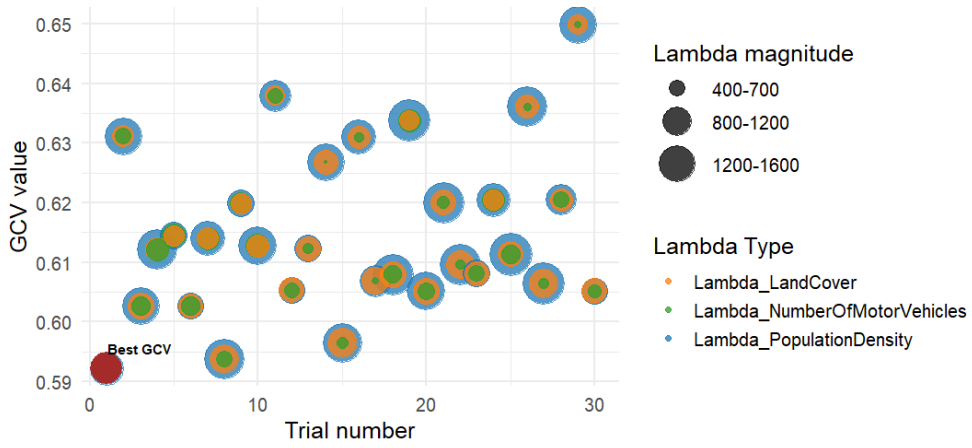

Figure 3. GCV values at different λ values

Figure 3 shows that the choice of the smoothing parameter substantially affects the shape of the estimated curve. Very small values of λtend to produce overly wiggly fits, whereas very large values lead to excessively smooth curves. The optimal $\lambda$was chosen as the value that minimized the GCV criterion.

2) Estimation Using the Smoothing Splines Method

The smoothing spline estimates using the optimal smoothing parameter for each predictor are presented in Figure 4. These values minimize overfitting and ensure accurate predictions, enhancing the model's ability to generalize effectively.

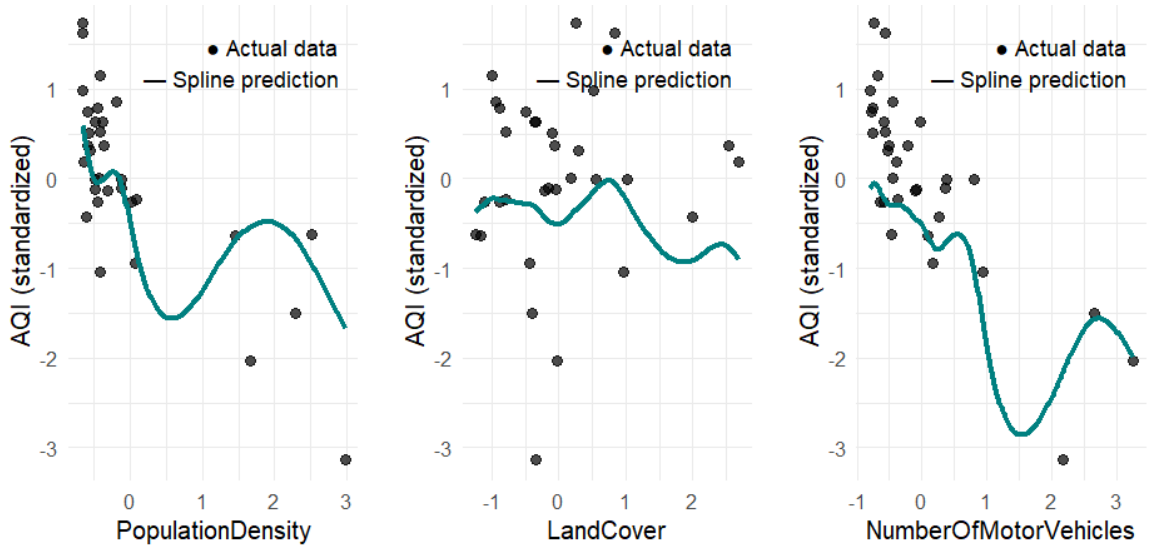

Figure 4. Smoothing Spline Predictions for Population Density, Land Cover, and Number of Motor Vehicles

Figure 4 illustrates the estimated smoothing spline curves for population density, land cover, and the number of motor vehicles. The fitted curves indicate that the effects of these predictors on AQI are nonlinear and vary across their ranges. Population density and the number of motor vehicles show more noticeable fluctuations, suggesting stronger changes in AQI across different levels of these variables. In contrast, the effect of land cover appears relatively more stable. These results support the use of smoothing splines for modeling AQI.

3) Model Evaluation of Smoothing Splines

Table 4. Model Evaluation Metrics Mean Squared Error and Adjusted R^2 of Smoothing Splines

Model Evaluation	Value
Mean Squared Error	0.0716
Adjusted R^2	0.7940

Based on Table 4 the Mean Squared Error (MSE) of the smoothing splines model, calculated from the residuals of each observation, measures the prediction error by squaring the differences between the observed and predicted AQI values. With an MSE of 0.0716, the model's predictions align closely with the observed data, indicating that the smoothing splines method accurately captures changes in air quality in the standardized dataset, reflecting the relationships between AQI, population density, land cover, and the number of motor vehicles. The low MSE suggests that

the model performs well even with regional differences and human activities. In addition, the Adjusted R^2 value of 0.794 indicates that approximately 79.4% of the variation in AQI can be explained by the predictor variables included in the model. These findings suggest that smoothing splines are effective in capturing the nonlinear relationship between AQI and the selected predictors.

Truncated Splines

After fitting the smoothing spline model, the analysis proceeded by estimating a truncated spline model to compare the performance of the two spline-based approaches. In the truncated spline model, candidate knot points were evaluated using the Generalized Cross Validation (GCV) criterion, and the best model was selected as the one with the minimum GCV value.

Table 5. Minimum GCV Values at Each Knots

Number of Knot Points	Minimum GCV Value
1 Knot Point	0.288
2 Knot Point	0.315
3 Knot Point	0.312

From Table 5 the minimum GCV values for each trial can be observed. To determine the best model, the lowest GCV value will be selected, which corresponds to the optimal knot points. Table 5 shows that the truncated spline model with one knot point produced the lowest GCV value (0.288). Therefore, this configuration was selected as the best truncated spline model for further evaluation.

1) Estimation of Nonparametric Truncated Splines Regression Model Parameters

Based on the evaluation results of all trials, the truncated splines model that provides the best performance is the model with one knot point. Next, the parameter estimation process is carried out using the Ordinary Least Squares (OLS) method to obtain the most accurate parameters. The truncated spline model with one knot point was then estimated using the Ordinary Least Squares (OLS) method, and its performance was evaluated using MSE and Adjusted R^2.

$$\hat{y} = 0.0348 - 0.291x_{i1} + 0.350(x_{i1} - 2.5208)_+ - 0.574x_{i2} - 13.678(x_{i2} - 2.204)_+ - 4.450x_{i3} + 3.387(xi_3 - 2.747)_+$$

2) Model Evaluation of Truncated Splines

Table 6. Model Evaluation Metrics Mean Squared Error and Adjusted R^2 of Truncated Splines

Model Evaluation	Value
Mean Squared Error	0.219
Adjusted R^2	0.729

According to Table 6, the truncated spline model yielded an MSE of 0.219 and an Adjusted R^2 of 0.729. These results indicate that the model was able to explain a substantial proportion of AQI variation, although its predictive performance was lower than that of the smoothing spline model.

Comparison of Smoothing Splines and Truncated Splines Performance

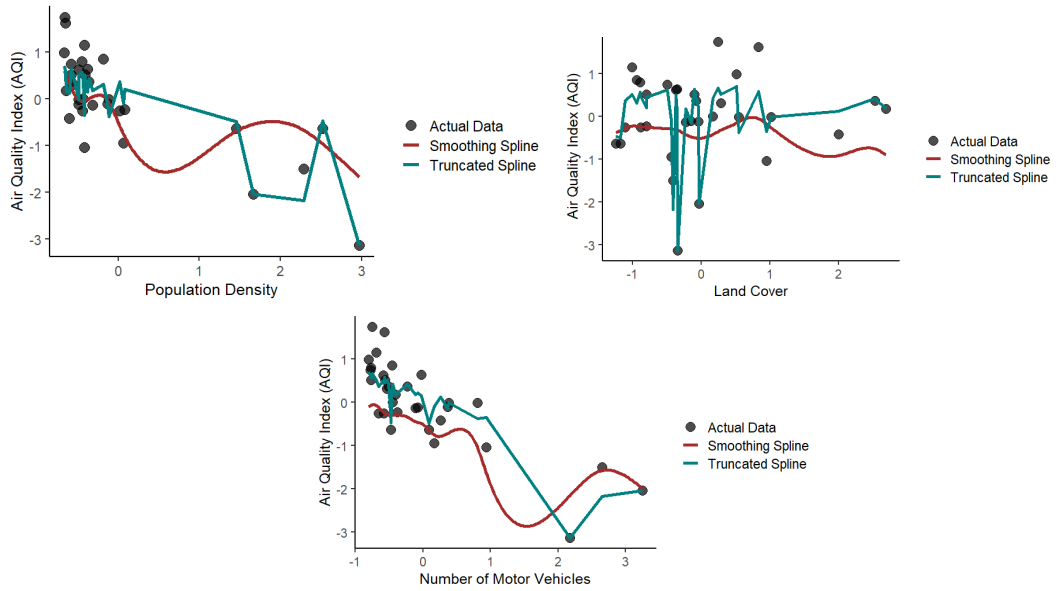

Figure 5. Comparison of Smoothing Splines and Truncated Splines Performance Across Different Predictors

Figure 5 compares the fitted curves from the smoothing spline and truncated spline models across the predictor variables. The smoothing spline curves appear more continuous and stable, whereas the truncated spline curves show sharper changes around the knot locations. This indicates that smoothing splines provide a more flexible and smoother representation of the nonlinear relationships in the data.

Table 7. Comparison of Model Performance Between Smoothing Splines and Truncated Splines

Model	Adjusted R^2	MSE
Smoothing Splines	0.794	0.0716
Truncated Splines	0.729	0.219

Based on Table 7, the smoothing spline model outperformed the truncated spline model in predicting AQI. This is indicated by its higher Adjusted R^2 value (0.794) and lower MSE value (0.0716), compared with the truncated spline model, which produced an Adjusted R^2 of 0.729 and an MSE of 0.219. These results indicate that smoothing splines are more effective in capturing the nonlinear relationship between AQI and the selected predictors.

CONCLUSION

This study compared smoothing splines and truncated splines for modeling the Air Quality Index (AQI) in Indonesia using provincial-level data. Both models were able to capture nonlinear relationships between AQI and the selected predictors. However, the smoothing spline model showed better performance, with a lower MSE (0.071) and a higher Adjusted R^2 (0.794) than the truncated spline model, which produced an MSE of 0.219 and an Adjusted R^2 of 0.729. These findings indicate that smoothing splines are more suitable for modeling AQI variation in Indonesia.

The results also suggest that population density and the number of motor vehicles have stronger nonlinear effects on AQI than land cover.

REFERENCE

Ariesta, D., Gusriani, N., & Parmikanti, K. (2021). Estimasi parameter model regresi nonparametrik B-splin E pada angka kematian maternal. *Jurnal Matematika UNAND*, *10*(3), 342–354. https://doi.org/10.25077/jmu.10.3.342-354.2021

Centofanti, F., Lepore, A., Menafoglio, A., Palumbo, B., & Vantini, S. (2023). Adaptive smoothing spline estimator for the function-on-function linear regression model. *Computational Statistics*, *38*(1), 191–216. https://doi.org/10.1007/s00180-022-01223-6

Chen, L., Smetanina, E., & Wu, W. B. (2021). Estimation of nonstationary nonparametric regression model with multiplicative structure. *The Econometrics Journal*, *25*(1), 176–214. https://doi.org/10.1093/ectj/utab018

Dani, A. T. R., & Adrianingsih, N. Y. (2021). Pemodelan regresi nonparametrik dengan estimator spline truncated vs deret fourier. *Jambura Journal of Mathematics*, *3*(1), 26–36. https://doi.org/10.34312/jjom.v3i1.7713

Dani, A. T. R., Adrianingsih, N. Y., Ainurrochmah, A., & Sriningsih, R. (2021). Flexibility of nonparametric regression spline truncated on data without a specific pattern. *Jurnal Litbang Edusaintech*, *2*(1), 37–43. https://doi.org/10.51402/jle.v2i1.30

Fatmawati, F., Budiantara, I. N., & Lestari, B. (2019). Comparison of smoothing and truncated spline estimators in estimating blood pressure models. *International Journal of Innovation, Creativity and Change*, *5*(3), 1177–1199.

Handayani, T., Sifriyani, S., & Dani, A. T. R. (2024). Stunting prevalence modeling using nonparametric regression of quadratic splines. *Jurnal Varian*, *7*(2), 149–160. https://doi.org/10.30812/varian.v7i2.2916

Mariati, N. P. A. M., Budiantara, I. N., & Ratnasari, V. (2021). Smoothing spline estimator in nonparametric regression (Application: Poverty in Papua Province). *Proceedings of the 7th International Conference on Research, Implementation, and Education of Mathematics and Sciences (ICRIEMS 2020)*, 309–314. https://doi.org/10.2991/assehr.k.210305.044

Nisa, K., Herawati, N., & Setiawan, E. (2018). Analisis regresi nonnparametrik dengan teknik smoothing. *Seminar Dan Rapat Tahunan (Semirata)*, 1–20.

Nurcahayani, H., Budiantara, I. N., & Zain, I. (2019). Nonparametric truncated spline regression on modelling mean years schooling of regencies in Java. *AIP Conference Proceedings 2194*, 020073. https://doi.org/10.1063/1.5139805

Pant, A., Joshi, R. C., Sharma, S., & Pant, K. (2023). Predictive modeling for forecasting air quality index (AQI) using time series analysis. *Avicenna Journal of Environmental Health Engineering*, *10*(1), 38–43. https://doi.org/10.34172/ajehe.2023.5376

Sahani, M., Singh, H., Patel, P. K., & Singh, S. (2024). Evaluation of predictive models for air quality index prediction in an Indian urban area. *Journal of Indian Association for Environmental Management*, *44*(3), 31–40.

Sriliana, I., Budiantara, I. N., & Ratnasari, V. (2022). A truncated spline and local linear mixed estimator in nonparametric regression for longitudinal data and its application. *Symmetry*, *14*(12), 2687. https://doi.org/10.3390/sym14122687

Suparti, S., Prahutama, A., & Santoso, R. (2018). Mix local polynomial and spline truncated: the development of nonparametric regression model. *Journal of Physics: Conference Series*, *1025*, 012102. https://doi.org/10.1088/1742-6596/1025/1/012102

Wongkar, D. C., Ruliana, R., & S., M. F. (2023). Analisis regresi nonparametrik spline truncated untuk menganalisis faktor-faktor yang mempengaruhi tingkat pengangguran terbuka di

Provinsi Sulawesi Selatan. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, *5*(2), 55–63.

Xu, D., & Wang, Y. (2021). Low-rank approximation for smoothing spline via eigensystem truncation. *Stat*, *10*(1). https://doi.org/10.1002/sta4.355