BAREKENG: Journal of Mathematics and Its Applications
June 2024 Volume 18 Issue 2 Page 1201–1216
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss2pp1201-1216

MODELLING CRIME RATES IN INDONESIA USING TRUNCATED SPLINE ESTIMATOR

Muhammad Althof Juniar [1], Azzahra Fania[2], Diana Ulya[3], Rico Ramadhan[4], Nur Chamidah[5*]

[1,2,3,4]*Statistics Study Program, Faculty of Science and Technology, Universitas Airlangga*
[5]*Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga*
Jln. Dr. Ir. H. Soekarno, Surabaya, 60115, Indonesia

*Corresponding author's e-mail: * **nur-c@fst.unair.ac.id***

Article History:

Received: 8th January 2024
Revised: 10th February 2024
Accepted: 7th April 2024
Published: 1st June 2024

Keywords:

Crime rate;
Decency;
Estimator Spline Truncated;
Narcotics.

ABSTRACT

Criminal acts are actions that violate the law and can arise from various factors such as emotions, psychological pressure, and others. Crime rate is a number that indicates the level of crime vulnerability in a certain area at a certain time. Higher crime rates correspond to increased vulnerability in an area and vice versa. Among various forms of criminal acts, the number of criminal acts and narcotics crimes in Indonesia tends to increase in 2020 and 2021. The aim of the research is to identify the characteristics of crime rate data based on the number of decency and narcotics incidents in Indonesia using a nonparametric regression approach. This research uses a nonparametric regression method spline truncated, and linear regression as a comparison. It was found that West Papua Province has the highest crime rate, based on a comparison between the linear regression model and the truncated spline nonparametric regression model, it can be concluded that the best model is the truncated spline nonparametric regression model with a Generalized Cross Validation (GCV) of 2468.487 and a coefficient of determination of 0 .7389091, indicating that approximately 73% of the variability of the dependent variable can be explained by the independent variables included in the model.

How to cite this article:

N. Chamidah, A. Fania, D. Ulya, M. A. Juniar and R. Ramadhan., "MODELLING CRIME RATES IN INDONESIA USING TRUNCATED SPLINE ESTIMATOR," *BAREKENG: J. Math. & App.*, vol. 18, iss. 2, pp. 1201-1216, June, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Criminal acts are acts that violate the law and can arise from various factors such as emotions, anger, psychological pressure, and other non-economic factors. According to Numbeo's 2021 data, Indonesia secured the 65th position out of 137 countries on the crime index, earning a score of 45.93. Numbeo clarified that a score below 20 signifies a very low crime rate, while a score exceeding 80 suggests a significantly high crime rate in a country. [1].

The crime rate is a numerical measure that reflects the extent of susceptibility to crime in a specific location during a particular period. A higher crime rate suggests an elevated vulnerability to crime in that area, while a lower crime rate indicates a reduced vulnerability. Based on data summarized by the Central Statistics Agency (BPS), during the period 2018 to 2021, the level of risk of being exposed to crime every 100,000 inhabitants experienced a decline. In 2021, it was 90 which decreased from 94 in 2020 and 103 in 2019.

Criminal acts are categorized in various forms, namely theft, drug use, immoral acts, pickpocketing, mugging, mugging with sharp/fire weapons, physical violence, abuse, destruction of other people's property, murder, fraud, and corruption [2]. The number of crimes that occurred during 2021 also decreased from the previous year, namely 2020 [3]. However, among these forms of criminal acts, the number of crimes in the form of decency and narcotics in Indonesia tended to increase in 2020 and 2021. For decency cases in 2020, there were 6,872 incidents, and in 2021, there were 5,905 incidents, and for narcotics cases in 2020, there were 36,611 incidents, and in 2021, there were 36,954 incidents [4].

Previous studies have provided insights into various aspects of crime dynamics across different regions of Indonesia. In a study focused on factors influencing crime in South Sulawesi, it was found that the most dominant variable affecting crime rates directly was the population size [5]. Similarly, research conducted on criminal acts in Tangerang City highlighted those variables such as population size, population growth rate, and the number of impoverished residents significantly influenced crime rates in the area [6]. Additionally, research into crime percentages in East Java revealed that factors such as the percentage of poor people, the open unemployment rate, gross regional domestic product on the basis of constant prices per capita, human development index, and average years of schooling have a significant influence on the crime rate [7]. This highlights a gap in previous research, which this current study seeks to address by extending its scope to incorporate decency-related offenses as key independent variables.

This research presents a novel approach to understanding crime rates by extending its scope beyond a single region, instead examining crime rates across all 34 provinces of Indonesia. Departing from previous research, this study incorporates two additional variables: the number of incidents of decency crimes and the number of narcotic crimes. Furthermore, the study employs nonparametric regression, a technique employed in statistical analysis, becomes great attention to flexibility and only assumes that the function form is smooth. Thus, nonparametric regression has the advantage of requiring fewer assumptions than parametric regression [8]. Nonparametric regression has various estimators. One of which is *Spline Truncated* or Spline Cut. On the estimator *Spline Truncated*, there are polynomial intersection points on different segments that are joined together at certain knots. By broadening the scope and variables, the research aims to provide a more comprehensive understanding of crime rate in Indonesia.

2. RESEARCH METHODS

2.1 Data Description

The data in this study case consists of secondary data obtained from the year 2021 Central Statistics Agency data in the book of Statistik Kriminal 2022 [9]. The variables used in this research include the number of incidents of crimes against decency (x_1), the number of incidents of crimes against narcotics (x_2), and the crime rate or level of risk of being exposed to crime (y).

2.2 Parametric Regression

Parametric regression models are employed when there is a recognizable pattern in the distribution of data, whether it be linear, quadratic, or cubic. Utilizing a parametric regression model necessitates an adequate amount of historical data or other information sources that offer insights into the available data. Within the realm of parametric regression, there is a distinct presumption that the form of the regression curve is predetermined. Specifically, the equation for linear parametric regression can be expressed as **Equation (1)** **[10]**.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_r x_{ri} + \varepsilon_i, \qquad i = 1,2,..,n \tag{1}$$

In this context, the y variable is the response variable, while x_1, x_2,…, x_r are the predictor variables. ε is an independent random error that has a normal distribution with mean zero and variance σ^2. The parameters β_0, β_1, β_2,… β_r are unknown parameters.

2.3 Nonparametric Regression

Nonparametric regression is employed when the connection between the response variable y and the predictor variable x is not known, and no specific assumptions are made about the form of the regression function. In nonparametric regression, the function is simply assumed to be smooth or within a specific function space. Compared to a parametric regression model, an approach utilizing nonparametric regression offers greater flexibility in determining the shape of the regression function. **[11]**.

Paired data (x_i,y_i) with n observations can be expressed in nonparametric regression model as follows:

$$y_i = g(x_i) + \varepsilon_i \tag{2}$$

where $g(x_i)$ is a regression function that is assumed to be smooth, that is, a continuous and differentiable function, while ε_i is a random error with zero mean and variance σ^2. The regression function $g(x_i)$ can be estimated using several techniques, including kernel, local linear, local polynomial, truncated spline, penalized spline, and Fourier series. The regression function estimation technique that will be used in this research is least squares spline or Truncated Spline.

2.4 Linear Regression

Regression analysis is a method of analysis that enables researchers to scrutinize data and derive significant insights regarding the correlation between two variables **[12]**. The linear regression equation is shown in **Equation (3)** as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_i X_i + \varepsilon \tag{3}$$

with

Y	: dependent variable
X_i	: independent variable
ε	: error
$\beta_0, \beta_1, \ldots, \beta_i$	: parameter regression model

There are two types of tests used to test linear regression parameters, namely simultaneous testing and partial testing which aims to determine the level of significance of the independent variables. Simultaneous parameter testing refers to a collective testing of all parameters in the regression model. The purpose of this test is to assess the significance of the β parameters on the response variable by utilizing analysis of variance (ANOVA) as presented in **Table 1.**

Table 2. ANOVA Test

Source	Sum square	DF	Mean Square Error	F Value
Regression	$\sum_{i=1}^{n} (\hat{y} - \bar{Y})^2$	p	$\frac{SSR}{p}$	$F = \frac{MSR}{MSE}$
Eror	$\sum_{k=1}^{n} (y_i - \hat{y}_i)^2$	$n - (p + 1)$	$\frac{SSE}{n - p - 1}$	
Total	$SSR + SSE$	$n - 1$		

The testing hypothesis in the simultaneous test is as follows.

$H_0 : \beta_1 = \beta_2$
H_1 : there is at least 1 $\beta_i \neq 0$ with $i = 1,2$

$$F_{value} = \frac{MSR}{MSE} \tag{4}$$

Rejection area: Reject H_0 if $F_{value} > F_{(\alpha;p;n-(p-1)}$ or if $p - value < \alpha$. Another method of parameter testing is partial testing, which aims to evaluate the significance of the β parameters on the response variable specifically by utilizing the t-test statistic.

$$t = \frac{\hat{\beta}}{SE(\hat{\beta}_i)} \tag{5}$$

where $SE(\hat{\beta}_i)$ is the standard error of the coefficient $\hat{\beta}_i$. Rejection area: Reject H_0 if $|t_{value}| > t_{(\frac{\alpha}{2};n-p-1)}$ or if $p = value < \alpha$.

Apart from that, linear regression also requires testing residual assumptions to assess the suitability of a model. These residual assumptions include normally distributed residuals, independent residuals, and identical residuals. Testing the assumption of normal distribution of residuals is carried out to evaluate whether the distribution of residuals meets the normal assumption. To test the normal distribution, the method that can be used is the Kolmogorov-Smirnov test.

Testing Hypothesis:

H_0 : The model residuals are normally distributed
H_1 : The model residuals are not normally distributed

$$D_{value} = Sup|F_n(x) - F_0(x)| \tag{6}$$

with
$F_n(x)$: Sample cumulative distribution value
$F_0(x)$: Cumulative distribution value below (for normal distribution: $P(Z < z_i)$
Rejection area: reject H_0 if $D_{value} > D(\alpha; N)$ and if $p - value < \alpha$.

Then, the residual independence assumption test was carried out to detect the possibility of correlation between residuals using the Durbin-Watson test method. The residual independence assumption test is carried out to detect the possibility of a correlation between the residuals. One testing method that is often used is the Durbin-Watson test. The hypothesis that is the basis for this test is as follows:

H_0 : independent Residuals
H_1 : Residuals are not independent

$$d = \frac{\sum_{i=1}^{n}(e_t - e_{t-1})}{\sum_{i=1}^{n} e_t^2} \tag{7}$$

with
e_t : residual on the second observation
t : many experimental observations
The test criteria are presented in **Table 2** as follows.

Table 2. Durbin Watson Test

H_0	Decision	Indicator
There is no positive autocorrelation	Reject H_0	0 < d < dL 0 < d < 1,3325
There is no positive autocorrelation	No Decision	dL < d < dU 1,3325< d < 1,5805
There is no negative autocorrelation	Reject H_0	4-dL < d < 4 2,6675 < d < 4
There is no negative autocorrelation	No Decision	4 – dU < d < 4-dL 2,4195 < d < 2,6675
There is no autocorrelation, positive or negative	Failed to Reject H_0	dU < d < 4-dU 1,5805< d < 2,4195

Moreover, the assumption of identical residuals is tested to find out whether the residual variance shows homoscedasticity. The technique that is often used is the Glejser test [13], with the hypothesis proposed as

H_0 : Identical residuals

H_1 : Residuals are not identical

This test is carried out by performing regression on the resulting model absolute residuals using all available predictor variables x. If the predictor variable is significant with $p - value < \alpha$, it can be concluded that there is heteroscedasticity, or it can be interpreted that the residuals do not meet the identical assumption. Conversely, if the predictor variable is not significant, this indicates that the residuals are identical or homoscedastic.

One of the important requirements in creating a regression model with several predictor variables is that there are no cases of multicollinearity, which indicates that there is no correlation between one predictor variable and other predictor variables. According to [14], the detection of multicollinearity cases can be done by looking at the Variance Inflation Factor (VIF) value, which must be more than 10. The VIF value can be expressed in Equation (8) as follows:

$$VIF = \frac{1}{1 - R_j^2} \tag{8}$$

Where R_j^2 is the coefficient of determination between X_j and other predictor variables. The way to overcome multicollinearity is to remove predictor variables that do not show significance from the model.

2.5 Spline Truncated

Truncated spline regression is a flexible polynomial cut. In general, $f(x_i)$ is a regression curve approximated by a truncated spline function with knot points namely $K_1, K_2, \dots, K_r$ which are derived from the following equation.

$$y_i = \sum_{j=0}^{p} \beta_j x_i^j + \sum_{k=1}^{r} \beta_{p+k} (x_i - K_k)_+^p + \varepsilon_i, i = 1, 2, \dots, n \tag{9}$$

Function $(x_i - K_k)_+^p$ is a truncated function given by Equation (10) [10].

$$(x_i - K_k)_+^p = \begin{cases} (x_i - K_k)^p, x_i \geq K_k \\ 0 \qquad\quad , xi < K_k \end{cases} \tag{10}$$

Where the point $x_i - K_k$ is the knot point which describes the pattern of changes in function in different sub-intervals and the p value is the degree of the polynomial.

2.6 Optimal Knot Selection

When utilizing splines, key alterations in the function's behavior at various intervals are denoted by the knot points. The selection of the optimal knot point in spline regression frequently relies on the Generalized Cross Validation (GCV) approach. The primary emphasis in identifying the most suitable spline regression model revolves around determining the knot points that exhibit the best conformity to the data. GCV is recognized for its asymptotically optimal properties in contrast to alternative methods like Cross Validation (CV) and the Unbiased Risk (UBR) method [15]. The main concept of GCV is a modification of CV [16]. Identification of optimal knot points is carried out by considering the minimum GCV value. The GCV method is generally defined as follows.

$$GCV(K_1, K_2, \dots, K_r) = \frac{MSE(K_1, K_2, \dots, K_r)}{(n^{-1} trace\,[I - A(K_1, K_2, \dots, K_r])^2} \tag{11}$$

2.7 Measures of Model Goodness

In this study case, accuracy was assessed using two metrics which are Mean Squared Error (MSE) and the coefficient of determination (R^2). MSE quantifies the precision of the model's predicted values through the average squared error, making it a useful tool for comparing forecast accuracy across various forecasting

methods [17]. Meanwhile, the coefficient of determination (R^2) is used to describe how much variation can be explained in the model. MSE is expected to have the minimum possible value, while the coefficient of determination (R^2) will be better the closer the value is to 1 [18]. The formula for MSE is as follows:

$$MSE(K_1, K_2, \ldots., K_r) = n^{-1} \sum_{i=1}^{n} \left(y - \hat{f}(x_i)\right)^2 \tag{12}$$

Meanwhile, the formula for the coefficient of determination (R^2) is as follows:

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \tag{13}$$

2.8 Step Analysis

1. In accordance with the first objective, namely identifying the characteristics of crime rate data based on the number of incidents of decency and narcotics in Indonesia with the following steps:
   a. Inputting data on predictor variables and response variables.
   b. Calculate the mean, median, standard deviation, minimum value, and maximum value with descriptive statistics.
2. In accordance with the second objective, namely modeling the crime rate based on the number of incidents of decency and narcotics in Indonesia with a linear regression approach with the following steps:
   a. Conduct a correlation test between predictor variables and response variables
   b. Modeling response variables and predictor variables using linear regression.
   c. Conducting simultaneous parameter significance test.
   d. Conducting partial parameter significance test.
   e. Performing residual assumption testing, which includes normality assumption test, independent residual assumption test, and identical residual assumption test.
   f. Performing multicollinearity testing.
3. In accordance with the third objective, namely modeling the crime rate based on the number of incidents of decency and narcotics in Indonesia with the Spline Truncated nonparametric regression approach with the following steps:
   a. Selecting the optimal knot points of the Spline Truncated model.
   b. Calculating the value of Generalized Cross Validation (GCV(λ)).
   c. Calculating the coefficient of determination (R^2).
   d. Estimating the parameters of the Spline Truncated nonparametric regression model.
   e. Interpreting the best nonparametric Spline Truncated regression model.
   f. Perform simultaneous parameter significance test.
   g. Testing the significance of parameters individually.
   h. Comparing the results of linear regression analysis with Spline Truncated nonparametric regression.

3. RESULTS AND DISCUSSION

3.1 Data Characteristics

Based on research data on crime rates based on the number of incidents of decency and narcotics in Indonesia, data is obtained and presented in descriptive statistics as follows.

Table 3. Descriptive Statistics of Crime Rates in Indonesia

Descriptive Statistics	Value
N	34
Mean	135.9
Standard deviation	69.2
Minimum	15.0

Median	122.5
Maximum	289.0

A visualization of the crime rate for each province in Indonesia is presented in **Figure 1** as follows.

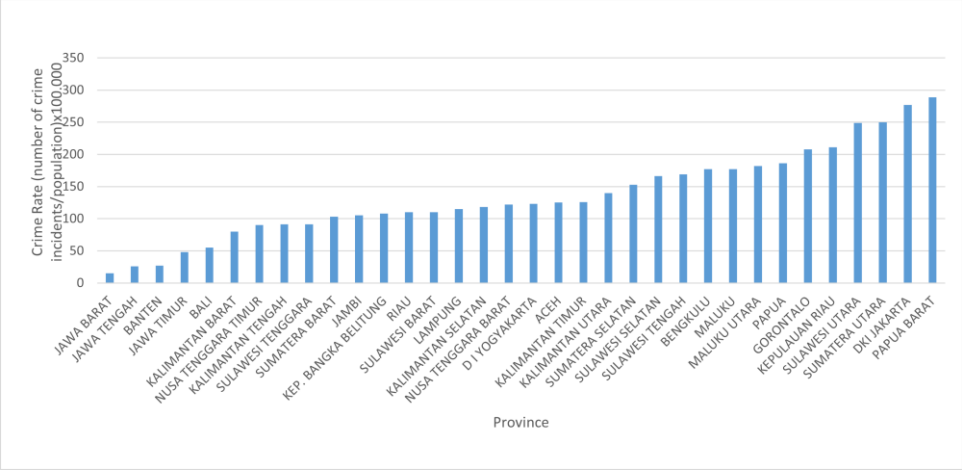


Figure 1. Graph of Crime Rates in Indonesia

Based on **Table 3** and **Figure 1** regarding the characteristics of the crime rate or level of risk of being exposed to crime, the highest crime rate occurred in West Papua Province at 289 (every 100,000 residents). The Crime Rate value of 289 in West Papua Province means that out of 100,000 residents in West Papua Province, 289 of them are victims of crime. Meanwhile, the lowest crime rate occurred in West Java Province at 15 (every 100,000 residents). The Crime Rate value of 15 in West Java Province means that out of 100,000 residents in West Java province, 15 of them are victims of crime. It can also be seen that the average (mean) crime rate is 135.9, with a standard deviation value of 69.2 and a median value of 122.5.

Table 4 shows the descriptive statistical values of the number of incidents of decency crimes that occurred in Indonesia in 2021.

Table 4. Descriptive Statistics on the Number of Incidents of Decency Crimes in Indonesia

Descriptive Statistics	Value
N	34
Mean	173.7
Standard Deviation	166.3
Minimum	20
Median	114
Maximum	904

Based on **Table 4** and **Figure 2** regarding the characteristics of the number of incidents of crimes against decency, North Sumatra Province is the region with the highest number of incidents of crimes against decency in Indonesia, namely 904 incidents. Meanwhile, North Kalimantan Province is the region with the fewest incidents of crimes against decency in Indonesia, namely 20 incidents. It can also be seen that the average (mean) number of incidents of decency crimes was 173.7 incidents, with a standard deviation value of 166.3 and a median value of 114.

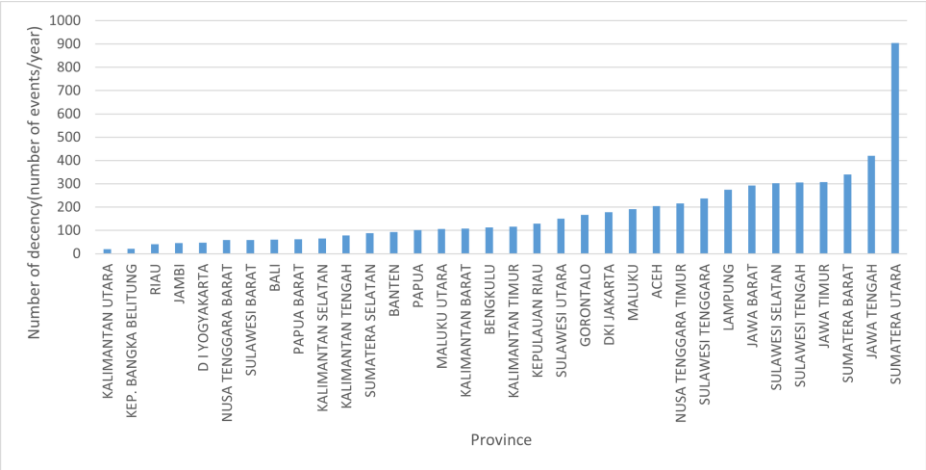


Figure 2. Graph of the Number of Incidents of Decency Crimes

Table 5 shows the descriptive statistical values of the number of narcotics crime incidents that occurred in Indonesia in 2021.

Table 5. Descriptive Statistics on the Number of Narcotics Crimes in Indonesia

Descriptive Statistics	Value
N	34
Mean	1087
Standard Deviation	1504
Minimum	2
Median	411
Maximum	5949

Based on **Table 5** and **Figure 3** regarding the characteristics of the number of narcotics crime incidents, North Sumatra Province is the region with the highest number of narcotics crime incidents in Indonesia, namely, 5949 incidents. Meanwhile, East Nusa Tenggara (NTT) Province is the region with the fewest incidents of narcotics crimes in Indonesia, namely, two incidents. It can also be seen that the average (mean) number of narcotics crime incidents is 1087 incidents, with a standard deviation value of 1504 and a median value of 411.

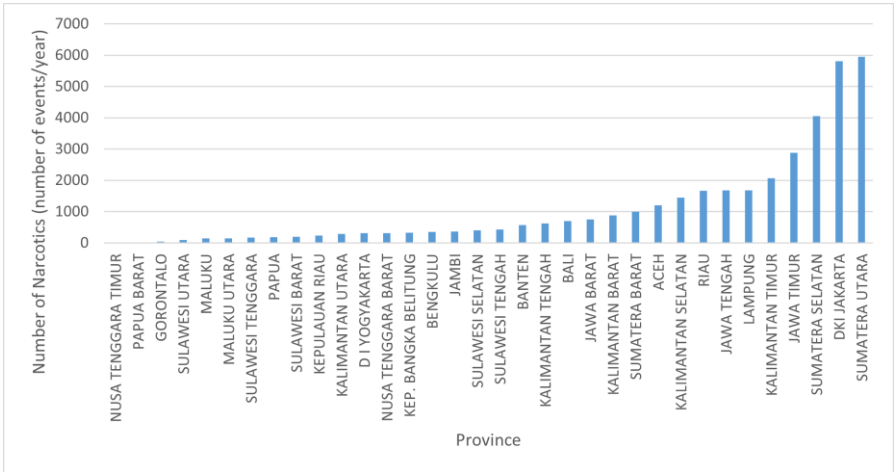


Figure 3. Graph of the Number of Narcotics Crime Incidents

3.2 Linear Regression

The first step in modeling data with linear regression is to check the correlation between the predictor variables and the response variable.

Table 6. Pearson Pairwise Correlation Test

Sample 1	Sample 2	Correlation	95% CI for ρ	P-Value
X_1	Y	0.133	(-0.215, 0.451)	0.455
X_2	Y	0.269	(-0.076, 0.557)	0.124

Table 6 shows that the $p - value > \alpha$ is between variables X_1 and Y and variables X_2 and Y so that a decision can be made to accept H_0. Therefore, it can be concluded that there is no linear relationship between the predictor variables and the response variables.

Linear regression analysis explores the linear connection between multiple independent variables and a dependent variable. The aim of this analysis is to ascertain whether a positive or negative correlation exists between the independent variable and the dependent variable, and also to forecast the dependent variable's value in response to changes in the independent variable, whether upward or downward. The linear regression equation formed is as follows.

$$Y = 117{,}5 - 0{,}0057X_1 + 0{,}01250X_2 \quad (14)$$

This equation can be interpreted as follows.

- A constant value of 117.5 means that if the variables for the number of incidents involving decency and narcotics are not included in the research, the crime rate will still increase by 117.5%.
- The coefficient value $\beta_1 = -0.0057$ means that if the variable number of crime incidents against decency increases then the crime rate will decrease by 0.0057% assuming the other independent variables are constant.
- The coefficient value $\beta_2 = 0.01250$ means that if the variable number of crime incidents involving narcotics increases, the crime rate will increase by 0.01250% assuming other independent variables are constant.

The concurrent examination aims to assess whether both the independent variable and the dependent variable exert an influence. The testing hypothesis in the simultaneous test is as follows.

$H_0 : \beta_1 = \beta_2$
H_1 : there is at least 1 $\beta_i \neq 0$ with $i = 1{,}2$

The test criteria are to reject H_0 if the $p - value > \alpha$ with a value of $\alpha = 0.05$. Simultaneous parameter significance results can be seen in **Table 7** as follows.

Table 7. Linear Regression ANOVA for Simultaneous Tests

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	11063	5531.5	1.21	0.311
X_1	1	8384	8383.63	1.84	0.185
X_2	1	21	21.24	0.00	0.946
Error	31	141439	4562.55		
Total	33	152502			

Table 7 shows that the $p - value > \alpha$ means that a decision can be made to accept H_0. Therefore, it can be concluded that simultaneously the regression model is not significant.

The partial parameter significance test in linear regression is used to assess the contribution of each independent variable to the dependent variable. This testing process involves hypothesis testing against each individual regression coefficient to determine whether the variable makes a significant contribution to explaining variation in the dependent variable. The hypothesis in this test is as follows.

$H_0 : \beta_i = 0, i = 1{,}2$

$H_1 : \beta_i \neq 0$

The test criteria are to reject H_0 if the $p - value < \alpha$ with a value of $\alpha = 0.05$. The partial parameter significance test results are presented in **Table 8** as follows.

Table 8. Partial Test Using T Test

Term	Coef	SE Coef	T-Value	P-Value
Constant	117,5	17.0	6.89	0.000
X1	-0.0057	0.0834	-0.07	0.946
X2	0.01250	0.00922	1.36	0.185

Table 8 shows that variables X_1 and X_2 have a $p - value > \alpha$ so that a decision can be made to accept H_0. Therefore, it can be concluded that partially, there are no variables that are significant or influence the crime rate.

The normality test is used to assess whether the distribution of residual values is normal or not. A good regression model is a model that has residual values that follow a normal distribution. The hypothesis in this test is as follows.

H_0: model residuals are normally distributed
H_1: model residuals are not normally distributed

The test criteria are to reject H_0 if the $p - value < \alpha$ with a value of $\alpha = 0.05$. The results of the normality test are presented in **Figure 4** as follows.

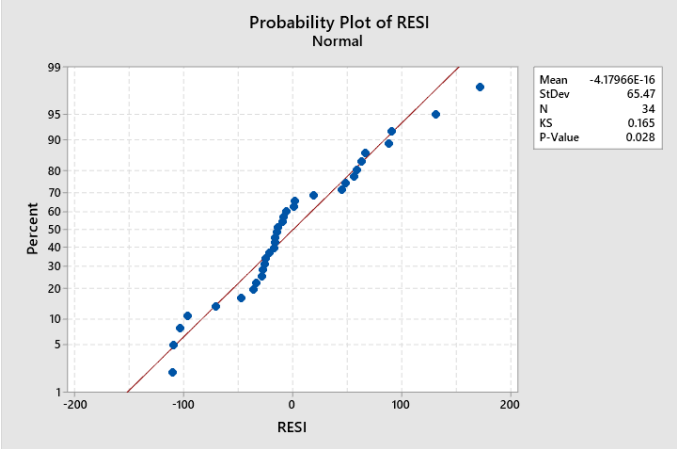


Figure 4. Residual Normality Test Graph (Source: Minitab 16)

Figure 4 shows that the $p - value < \alpha$ so that a decision can be made to reject H_0. Therefore it can be concluded that the model residuals are not normally distributed.

The purpose of the autocorrelation test is to establish if there is a correlation between the current period (t) and the preceding period (t – 1). The results of the autocorrelation test with the Durbin Watson test are presented in **Table 9** as follows.

Table 9. Autocorrelation Test with Durbin Watson

Test	Test Statistics
Durbin Watson	1,57026

Because the calculated d is 1.57026 and is in the range 1.3325 < d < 1.5805, it can be concluded that there is no decision regarding the autocorrelation test.

The heteroscedasticity test aims to determine whether there are differences in the variance of the residuals between observations. A regression model is considered eligible when the residual variance between observations remains constant, which is called homoscedasticity. The heteroscedasticity test uses the Glejser method by correlating the independent variables with their absolute residual values. The hypothesis in this test is as follows.

H_0: Identical residuals (no heteroscedasticity)
H_1: Residuals are not identical (heteroscedasticity occurs)
The test criteria are to reject H_0 if the $p - value < \alpha$ with a value of $\alpha = 0.05$. The results of the heteroscedasticity test are presented in **Table 10** as follows.

Table 10. Glejser Test Results Using T-Test

Term	Coef	SE Coef	T-Value	P-Value
Constant	-0.0	17.0	-0.00	1.00
X1	-0.0000	0.0834	-0.00	1.00
X2	0.00000	0.00922	0.00	1.00

Table 10 shows that the $p - value > \alpha$ means that a decision can be made to accept H_0. Hence, one can infer that the residuals are either identical or there is an absence of heteroscedasticity.

This test aims to reflect the existence of a close relationship in linear form between several predictor variables in a simple linear regression model. A good regression model should involve predictor variables that are independent or not correlated with each other. In testing this assumption, it is desired that the multicollinearity assumption is not met. **Table 11** shows the results of the multicollinearity test as follows.

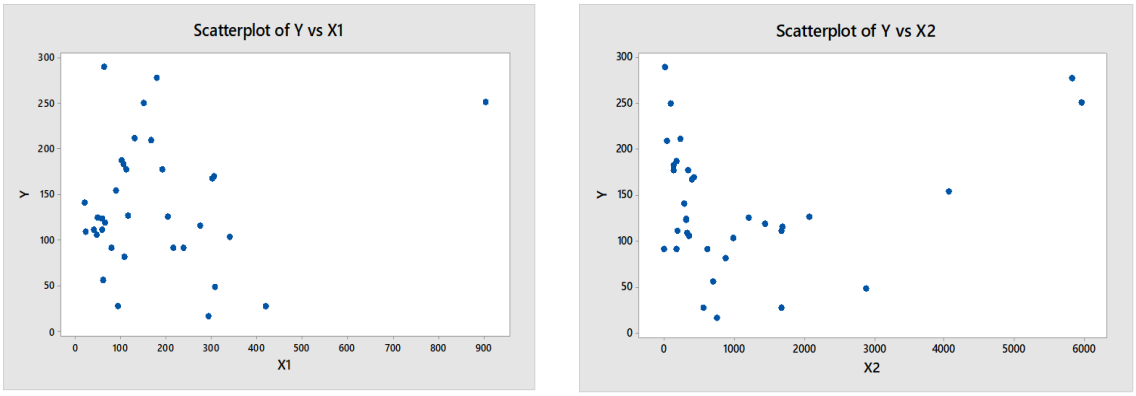
Table 11. Multicollinearity Test Results

Term	VIF
X_1	1.39
X_2	1.39

The presence of multicollinearity can be assessed through the Variance Inflation Factor (VIF) column. If the values in the VIF column remain below 10, it indicates the absence of multicollinearity in the independent variable. In other words, there is no substantial relationship with the independent variable.

3.3 Nonparametrics Regression Truncated Spline

The first step, before modeling with truncated spline, is identify the relationship pattern of the response variable, namely the crime rate with the predictor variable, by showing scatterplot.



(a) (b)

Figure 5. Scatterplot, (a) Variable Number of Decency Incident, (b) Variable Number of Narcotics Incident (Source: Minitab 16)

From **Figure 5**, it can be seen that the relationship patterns formed visually do not form a particular pattern. Therefore, model estimation cannot be done using a parametric regression approach so that the variables are non-parametric components. So, non-parametric regression is needed to determine the influence of the variable number of incidents of decency crimes and the variable number of incidents of narcotics crimes on the crime rate in Indonesia.

The best truncated spline nonparametric regression model is obtained from optimal knot points. To obtain the optimal knot point, the minimum and maximum GCV values are used. The following is the selection of optimal knot points with one knot point.

Table 12. Selection of Optimal Knot Points with One Knot Point

No	x_1	x_2	GCV	R^2
1	39	124	5033.46	25.64181
2	57	245	4336.88	35.93223
3	75	367	4076.848	39.77363
4	93	488	3458.852	48.90314
5	111	609	3137.81	53.64582
6	**129**	**731**	**3097.348**	**54.24356**
7	147	852	3260.101	51.83924

Based on Table 12, it can be seen that the minimum GCV value is 3097.348 and the maximum R^2 is R^2 = 54.2%. The optimal knot points from the minimum GCV and maximum R^2 are as follows:

$$K_1 = 129; \quad K_2 = 731$$

These results will be compared using two knots and three knots. This comparison was carried out to get the best Spline model results. The following is the selection of optimal knot points with two knot points.

Table 13. Selection of Optimal Knot Points with Two Knot Points

No	x_1	x_2	GCV	R^2
1	129 731	868 5707	3200.131	59.7185
2	129 731	886 5828	3110.146	57.5208
3	**129** **731**	**904** **5949**	**3097.312**	**56.24354**

Based on Table 13, it can be seen that the minimum GCV value is 3097.312 and the maximum R^2 is R^2 = 56.2%. The optimal knot points from the minimum GCV and maximum R^2 values are as follows:

$$(K_1 = 129;\ K_2 = 731), \quad (K_3 = 904;\ K_4 = 5949)$$

The following is the selection of optimal knot points with three knot points.

Table 14. Selection of Optimal Knot Points with Three Knot Points

No	x_1	x_2	GCV	R^2
1	147 273 1338	219 852 1702	2583.519	72.67422
2	147 291 1338	219 852 1823	2652.235	71.94742
3	**147** **291** **1216**	**201** **852** **1823**	**2468.487**	**73.89091**

Based on Table 14, it can be seen that the minimum GCV value is 2468.487 and the maximum R^2 is R^2 = 73.89%. The optimal knot points from the minimum GCV and maximum R^2 values are as follows:

$$(K_1 = 147;\ K_2 = 291;\ K_3 = 1216),$$

$$(K_4 = 201;\ K_5 = 852;\ K_6 = 1823)$$

Based on the previous discussion, namely spline modeling with one, two, and three knot points, a comparison of these models was carried out to determine the best model. It turns out that the minimum GCV and maximum R^2 values are the GCV and R^2 values with three knot points. As seen in Table 15 below.

Table 15. Comparison of GCV and R^2 Values for Various Knots

Knot	GCV	R^2
One	3097.312	56.24356
Two	3097.312	56.24356
Three	**2468.487**	**73.89091**

The outcomes of estimating parameters with three knot points are as stated.

$$\hat{y} = 174{,}99 + 0{,}73x_1 - 4{,}09(x_1 - 147)_+ + 4{,}8(x_1 - 201)_+ \\ -1{,}56(x_1 - 291)_+ - 0{,}25x_{2_+} + 0{,}88(x_2 - 852)_+ \\ 0{,}97(x_2 - 1216)_+ + 0{,}4(x_2 - 1823)_+ \quad (15)$$

The best model can be interpreted as follows.

1. If the variable is considered constant, then the number of cases of decency relative to the crime rate is

$$\hat{y} = 174{,}99 + 0{,}73x_1 - 4{,}09(x_1 - 147)_+ + 4{,}8(x_1 - 201)_+ \\ 1{,}56(x_1 - 291)_+ - 0{,}25x_{2_+} + 0{,}88(x_2 - 852)_+ \\ 0{,}97(x_2 - 1216)_+ + 0{,}4(x_2 - 1823)_+ \quad (16)$$

Based on this equation, it can be explained as follows.

$$\hat{y} = \begin{cases} 174{,}99 + 0{,}73x_1, & x_1 < 147 \\ 773{,}27 - 3{,}36x_1, & 147 \le x_1 < 201 \\ -188.645 + 1{,}44x_1, & 201 \le x_1 < 291 \\ 264{,}7 - 0{,}12x_1, & x_1 \ge 291 \end{cases} \quad (17)$$

In cases of decency in Indonesia, if the number is below 147 cases, then for every increase in one case of decency, the crime rate increases by 0.73. In cases of decency, with a number between 147 and under 201 cases, for every increase of one case, the crime rate decreases by 3.36. In decency cases with a number between 201 and under 291 cases, for every increase of one case, the crime rate increases by 1.44. And 291 cases of decency or more than that for every increase of one case have an effect on decreasing the crime rate by 0.12.

2. If the variable is considered constant, then the number of narcotics cases will be greater than the crime rate

$$\hat{y} = 174{,}99 - 0{,}25x_{2_+} + 0{,}88(x_2 - 852)_+ \\ -0{,}97(x_2 - 1216)_+ + 0{,}4(x_2 - 1823)_+ \quad (18)$$

Based on this equation, it can be explained as follows.

$$\hat{y} = \begin{cases} 174{,}99 - 0{,}25x_2, & x_2 < 852 \\ -574{,}77 + 0{,}63x_2, & 852 \le x_2 < 1216 \\ 604{,}75 - 0{,}34x_2, & 1216 \le x_2 < 1823 \\ -124{,}45 + 0{,}06x_2, & x_2 \ge 1823 \end{cases} \quad (19)$$

In narcotics cases in Indonesia, if the number is below 852 cases, then for every increase in one narcotics case, the crime rate decreases by 0.25. In the number of narcotics cases between 852 and under 1216 cases, for every increase of one case, the crime rate increases by 0.63. In the number of narcotics cases between 1216 and under 1823 cases, for every increase of one case, the crime rate decreases by 0.34. And there were 1,823 narcotics cases or more. For every increase of one case, the crime rate increased by 0.06.

To find out whether the parameters obtained from modeling results using truncated spline nonparametric regression have a significant influence on the crime rate in Indonesia, parameter significance testing was carried out.

Table 16. ANOVA Table for Spline Regression

Source	Df	*Sum of Square*	*Mean Square*	F_{test}	*P-value*
Regression	8	109072.1	13.634,02	7.782	0,.00061
Error	22	38540.26	1.751,83		
Total	30	147612.4			

According to the ANOVA findings, the P-value is observed to be 0.00061. This value is less than the value α = 0.05. So, in this case, it can be concluded that if it is rejected, there is at least one parameter that is significant to the response variable. To be able to find out which parameters influence the response variable, it is necessary to carry out individual testing.

Table 17. Spline Regression Parameters Individually Test

Variable	Parameter	Coefficient	*P-value*	Decision
x_1	β_1	174.99	0.000003	Significant
	β_2	0.73	0.02	Significant
	β_3	-4.09	0.0013	Significant
	β_4	4.8	0.0025	Significant
	β_5	-1.56	0.019	Significant
x_2	β_6	-0.25	0.000017	Significant
	β_7	0.88	0.00016	Significant
	β_8	-0.97	0.00097	Significant
	β_9	0.4	0.00167	Significant

According to the provided table, it is evident that every parameter holds significance. This parameter is significant at the 0.05 significance level because p-value < α. Overall, the parameters are thought to influence the crime rate. After testing the data on parametric analysis with linear regression and non-parametric with truncated splines, the results of the two tests are summarized in **Table 18** below:

Table 18. Comparison between Parametric and Nonparametric test

ANALYSIS	MSE	R^2
Linear Regression	4562.55	7.25%
Spline Truncated	1751.83	73.89%

The optimal model chosen for the two predictor variables was determined by comparing the MSE value to the reference value specified by the author. The preferred model turned out to be a Truncated Spline nonparametric regression model with a GCV of 2468.487.

4. CONCLUSIONS

Following the examination and discussion results, the derived conclusions are as follows.

1. The highest crime rate occurred in West Papua Province at 289 (every 100,000 residents), meaning that out of 100,000 residents in West Papua Province, 289 of them were victims of crime. North Sumatra Province is the region with the most incidents of crimes against decency and narcotics in Indonesia, namely 904 incidents and 5949 incidents.
2. By using linear regression, an equation is formed with a value of 7.25% as follows.

$$Y = 117{,}5 - 0{,}005X_1 + 0{,}01250X_2$$

3. From the results of selecting optimal knot points, the regression model using three knot points is the best, with a GCV value of 2468.487 and 0.7389091 or 73.89%, and the following modeling form.

$$\hat{y} = 174{,}99 + 0{,}73x_1 - 4{,}09(x_1 - 147)_+ + 4{,}8(x_1 - 201)_+ - 1{,}56(x_1 - 291)_+ - 0{,}25x_{2+} + 0{,}88(x_2 - 852)_+ - 0{,}97(x_2 - 1216)_+ + 0{,}4(x_2 - 1823)_+$$

REFERENCES

[1] Numbeo, "Crime Index by City 2024", 2024, [Online], Tersedia: https://www.numbeo.com/crime/rankings.jsp [Diakses: 20 Oktober 2023].

[2] Wilona, M., "Representasi Kriminalitas dalam Film Ted dan Ted 2," Jurnal E-Komunikasi, vol. 3, no. 2, pp. 1-12, 2015.

[3] Badan Pusat Statistik Indonesia, "Statistik Kriminal 2021", 2022, [Online], Tersedia: https://www.bps.go.id/id/publication/2021/12/15/8d1bc84d2055e99feed39986/statistik-kriminal-2021.html [Diakses pada 16 September 2023].

[4] Syaputra, H., "Tingkat Kejahatan Selama Pandemi", 2022, [Online], Tersedia: https://news.detik.com/kolom/d-5926380/tingkat-kejahatan-selamapandemi [Diakses pada 16 September 2023].

[5] Zaki, A., and Ilham, M., "Pemodelan Jalur pada Faktor yang Mempengaruhi Kriminalitas di Sulawesi Selatan Tahun 2021," Jurnal Matematika dan Statistika serta Aplikasinya, vol. 10, no. 1, pp. 1-8, 2022.

[6] Ningsih, D. R., Intan, P. K., and Yuliati, D., "Pemodelan Tindak Pidana Kriminalitas di Kota Tangerang Menggunakan Metode Regresi Lasso," ESTIMASI: Journal of Statistics and Its Application, pp. 64-77, 2023.

[7] Febrianti, E., Susetyo, B., and Silvianti, P., "Pemodelan Tingkat Kriminalitas di Indonesia Menggunakan Analisis Geographically Weighted Panel Regression," Xplore: Journal of Statistics, vol. 12, no. 1, pp. 91–109, 2023.

[8] Nurdiani, N., Herrhyanto, N., and Dasari, D., "Regresi Nonparametrik Birespon," EurekaMatika, vol. 5, no. 1, pp. 106-121, 2017.

[9] Badan Pusat Statistik Indonesia, "Statistik Kriminal 2022," 2023, [Online], Tersedia: https://www.bps.go.id/id/publication/2022/11/30/4022d3351bf3a05aa6198065/statistik-kriminal-2022.html [Diakses pada 16 September 2023].

[10] Chamidah, N., and Lestari, B., "Analisis Regresi Nonparametrik dengan Perangkat Lunak R," Airlangga University Press, 2022.

[11] Prasetyo, R. A., and Helma, "Analisis Regresi Linear Berganda Untuk Melihat Faktor yang Berpengaruh Terhadap Kemiskinan di Provinsi Sumatera Barat," *Journal Of Mathematics UNP*, vol. 7, no. 2, pp. 62-68, 2022.

[12] Sulistyono, S. and Sulistiyowati, W., "Peramalan Produksi dengan Metode Regresi Linier Berganda," PROZIMA (Productivity, Optimization and Manufacturing System Engineering), vol. 1, pp. 82, 2018, https://doi.org/10.21070/prozima.v1i2.1350

[13] R. B. Darlington and A. F. Hayes, "Regression analysis and linear models: Concepts, applications, and implementation," Guilford Publications, 2016.

[14] J. Frost, "Regression Analysis: An Intuitive Guide for Using and Interpreting Linear Models," Statistics by Jim Publishing, 2019.

[15] S. Hartshorn, "Linear Regression And Correlation: A Beginner's Guide," Amazon digital services LLC, 2017.

[16] Pratiwi, L.P.S., Ayuningsih, N.P.M., and Dwijayani, N.M, "Perbandingan GCV dan UBR dalam Regresi Nonparametrik Multivariabel," Jurnal Matematika, vol. 11, no. 1, pp. 64-74, 2021.

[17] Pratiwi, L.P.S., Ayuningsih, N.P.M., and Wijaya, I.M.P.P., "Perbandingan Metode CV dan GCV pada Pemodelan MARS (Aplikasi Rata-Rata Lama Sekolah di Kabupaten Gianyar)," SAINTIFIK: Jurnal Matematika, Sains dan Pembelajarannya, vol. 8, no. 2, pp. 114-122, 2022.

[18] Abdullah, S.M., and Novianti, W. "Perancangan Sistem Informasi Peramalan Penjualan Meubel menggunakan Metode *Moving Average* (Studi Kasus Toko Meubel Sumber Rejeki)," INFORMASI INTERAKTIF: Jurnal Informatika dan Teknologi Informasi, vol. 7, no. 2, pp. 96-100, 2022.