Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 17(2) 2024: 221-232
http://ejournal.undip.ac.id/index.php/media_statistika

THE ANALYSIS OF SOCIO-ECONOMIC EFFECT ON CRIMINALITY IN INDONESIA USING FUZZY CLUSTERWISE REGRESSION MODEL

Dian Fatimah Azzarah, Moch. Abdul Mukid, Di Asih I Maruddani, Masithoh Yessi Rochayani
Department of Statistics, Diponegoro University, Semarang, Indonesia

e-mail: mamukid@live.undip.ac.id

DOI: 10.14710/medstat.17.2.221-232

Article Info:
Received: 15 September 2024
Accepted: 14 September 2025
Available Online: 11 October 2025

Keywords:
Crime; Socio-economic Factors; Fuzzy Clusterwise Regression
.

Abstract: Crime in Indonesia has shown a fluctuating trend and has increased significantly in recent years, with striking variations in crime rates between provinces. This phenomenon raises questions about the role of socio-economic factors such as education, poverty, and unemployment in influencing crime rates. Although there have been many studies examining the relationship between these variables and crime, the approaches used often assume that the relationship between variables is homogeneous across regions. In fact, heterogeneity in characteristics between provinces can cause different relationships. Therefore, an analysis approach is needed that can accommodate this diversity. This study proposes the Fuzzy Clusterwise Regression method which not only improves model accuracy compared to classical linear regression (with an increase in the coefficient of determination from 65.72% to more than 90%), but is also able to identify different patterns of relationships between regional groups (clusters). The results from FCR showed that the effect of socio-economic factors on crime varies between clusters and the optimum number of clusters is 4. In cluster 1, cluster 2, and cluster 3 all the variables had a significant influence on the amount of crime. Meanwhile, in cluster 4, the population poverty variable has no significant effect on the crime rate.

1. INTRODUCTION

Crime is a form of social action that violates legal norms related to seizing the property of others, disturbing public order, and killing one or a group of people (Kartono, 2009). Criminal acts occur because of social inequality, hatred, mental stress, or environmental changes that occur in society (Soekanto et al., 1986). Crime has a broad impact on all levels of society, crimes often occur in various places and at different times (Wahyuni P., 2010). The emergence of various types of crimes shows that criminality is always developing (Setiadi, 2000). (Goulas & Zervoyianni, 2015) state that crime is relatively harmless if followed by satisfactory economic conditions. Satisfactory conditions are when the employment ratio of the population is above average and life expectancy is improving. According to Grover (2012), socio-economic status drives criminal behavior, more than 67% of prison inmates in the UK are unemployed, and 48% of prisoners have a history of debt that they are unable to pay off. Socio-economic factors are important factors that influence the crime rate in a society (Breetzke & Pearson, 2015).

Many socio-economic factors influence crime. Education is one of them. The level of education is expected to reduce criminal behavior by increasing the chances of getting legal employment. (O'Sullivan, 2019) stated that college graduates earn at least twice as much as high school graduates and high school graduates earn almost 1.5 times more than those who drop out of school. According to research by Arsono & Atmanti (2014), the low level of education a person has will result in increasingly narrow opportunities to enter the labor market and the increasing difficulty of increasing work productivity, which will have an impact on the high number of unemployed. Lochner & Moretti (2004) also argue that the lower a person's level of education, the lower their skills are compared to high school to university graduates, and the free time that elementary school to high school graduates have will be more than high school to university graduates. So that the availability of excess free time can be an opportunity for them to commit crimes. (Lochner, 2020) study also stated that an increase in schooling significantly decreases the risk of violent and property crimes. (Mohammed & Mohamed, 2015) found that prisoners who participated in educational programs while in prison were less likely to re-offend than those who did not participate in skills education programs. Therefore, prisoners who are apathetic towards educational programs while in prison tend to repeat criminal acts. Other socio-economic factors that drive crime are poverty. The chain of poverty will trigger various problems such as unemployment, hunger, ignorance, crime, and others. A person's inability to meet their needs can trigger theft, murder, fraud, robbery, and so on (Pare & Felson, 2014).

Unemployment is also a socio-economic factor that drives crime. Unemployment increases poverty because people without jobs have no income to meet their basic needs. The lack of employment opportunities forces some individuals to seek illegal ways to earn money. As a result, economic hardship can drive people toward criminal behavior. According to Melick (2003), unemployed individuals are more likely to commit crimes because they have no legal source of income.

According to data from the Indonesian Central Statistics Agency (BPS, 2023), the number of crimes in Indonesia experienced quite significant fluctuations in the period 2020 to 2022. In 2020, the number of crimes recorded was 247,218 cases. This figure then decreased slightly in 2021 to 239,481 cases. However, there was a fairly drastic spike in 2022, when the number of crimes reached 372,965 cases. This sharp increase indicates an increase in criminal activity in Indonesia that year. Crime data in Indonesia shows quite significant disparities between provinces. East Java is in the top position with the highest number of crime cases, followed by North Sumatra and Metro Jaya. Provinces with large economic centers and dense populations tend to have higher crime rates. Conversely, provinces in eastern Indonesia such as Papua, Maluku, and East Nusa Tenggara generally have lower crime rates.

Various studies have shown that socioeconomic factors have a significant influence on crime rates in a society. Individuals with low levels of education tend to have fewer job opportunities, which ultimately increases the risk of involvement in criminal activities (Lochner & Moretti, 2004). In addition, poverty is often associated with less stable environmental conditions and social disorganization, which trigger high crime rates, especially property and violent crimes (Pare & Felson, 2014). Unemployment is also a major trigger, because individuals who do not have a steady income are more likely to commit crimes to meet their living needs (Melick, 2003). Research in the UK shows that more than 67% of prisoners were unemployed before being imprisoned, confirming the link between economic instability and crime (Grover, 2012). Therefore, improving the quality of education, reducing poverty rates, and creating jobs are important strategies in reducing crime rates.

This disparity indicates that there are complex factors that influence crime rates in each region. These factors can include the level of urbanization, socio-economic conditions, education levels, employment opportunities, and the effectiveness of law enforcement. In addition, the types of crimes that are dominant in each province can also vary, so further analysis is needed to understand the causes behind the differences in crime rates between regions. One of the statistical tools that can be used to study this phenomenon is by using a clusterwise regression model. This method allows the identification of groups or clusters in the data, each of which has a different regression model, thus capturing local variations and producing more accurate models than classical linear regression which assumes a homogeneous relationship. In addition, clusterwise regression is useful for uncovering hidden patterns that are not detected in traditional approaches.

Clusterwise regression models are valuable analytical tools for identifying heterogeneity in data (Desarbo & Cron, 1988). In many cases, the assumption that the relationship between predictor variables and response variables is linear and constant across the data is not always valid. Clusterwise regression models address this by identifying groups of data (clusters) that have different relationship characteristics. In doing so, we can gain a deeper understanding of the data and build more accurate models for each group. The importance of the clusterwise regression model lies in its ability to reveal hidden patterns in complex data. This model allows us to identify subgroups within a population that may have different responses to predictor variables. This is very useful in various fields, such as marketing, health, and social sciences. For example, in marketing, this model can be used to identify different customer segments and develop more effective marketing strategies for each segment.

Based on the description above, this research was conducted to further analyze the relationship between education, poverty, and unemployment with crime in Indonesia using the Fuzzy Clusterwise Regression method to see the linear relationship between response variables and predictor variables and detect clusters involving different relationships.

2. LITERATURE REVIEW

The first person who developed clusterwise regression was Oldenburg H. Spath (Spath, 1979). The main weakness of the Spath algorithm is its heuristic nature and its high dependence on the initial partition chosen. This algorithm uses an exchange method to move observations between clusters in order to minimize the sum of squared errors within the cluster. However, the final result is not guaranteed to be optimal because the algorithm only searches for a "good" local solution, not the best global solution. Fuzzy Clusterwise Linear Regression (FCR) is an alternative method developed to overcome the weaknesses of the conventional Clusterwise Linear Regression method. FCR is a method that combines fuzzy clustering and regression to identify correlations between response variables and predictor variables, and group data into a number of clusters with distinct relationships (Wedel & Steenkamp, 1989). Assume that the parameter vectors differ across clusters, and the number of clusters is known. Each cluster is assumed to have a different parameter vector of size $(\mathrm{P}+1) \times 1$. The general model for Fuzzy Clusterwise Linear Regression is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, 2, \dots, c \qquad (1)$$

Where $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1P} \\ 1 & x_{21} & x_{22} & \dots & x_{2P} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{N1} & x_{N1} & \dots & x_{nP} \end{pmatrix}$, $\boldsymbol{\beta}_\mathbf{i} = \begin{pmatrix} \beta_{i0} \\ \beta_{i1} \\ \vdots \\ \beta_{iP} \end{pmatrix}$, $\boldsymbol{\varepsilon}_{\boldsymbol{i}} = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{in} \end{pmatrix}$

y is a response vector $(n \times 1)$, **X** is a predictor matrix $(n \times (P+1))$, $\boldsymbol{\beta}_i$ is the parameter vector of the i-th cluster $((P+1) \times 1)$, and $\boldsymbol{\varepsilon}_i$ is a vector of residuals of the i-th cluster $(n \times 1)$. *P* is the number of independent variables, *n* is the number of observation units, *c* is the number of clusters, and *m* is the fuzzy parameter.

The objective function of FCR is to minimize

$$F = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} \varepsilon_{ij}^{2} \tag{2}$$

subject to $\sum_{i=1}^{c} u_{ij} = 1$, and $0 \le u_{ij} \le 1$, where u_{ij} is the fuzzy membership of the j-th object $(j = 1, 2, \dots, n)$ of the i-th cluster $(i = 1, 2, \dots, c)$. m is a fuzzy parameter and commonly m = 2 is used because this value is considered the smoothest (Klawoon & Hoppner, 2003; Wu, 2012). The constrained optimization problem stated in Equation (2) can be transformed into a non-constrained optimization problem using the Lagrangian Multiplier (Bertsekas, 1982),

$$J = F + \sum_{j}^{n} \lambda_j \left(\sum_{i=1}^{c} u_{ij} - 1 \right) \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} \varepsilon_{ij}^{2} + \sum_{j}^{n} \lambda_j \left(\sum_{i=1}^{c} u_{ij} - 1 \right) \tag{3}$$

where λ_j is a Lagrangian parameter.

In order to determine the estimator for $\boldsymbol{\beta}_\mathbf{i}$, Equation (3) is transformed into

$$J = \sum_{i=1}^{c} J_i + \sum_{j}^{n} \lambda_j \left(\sum_{i=1}^{c} u_{ij} - 1 \right) \tag{4}$$

where $$J_i = \sum_{j=1}^{n} u_{ij}^{m} \varepsilon_{ij}^{2}\, u_{i1}{}^{m}\, \varepsilon_{i1}{}^{2} + u_{i2}{}^{m} \varepsilon_{i2}{}^{2} + \dots + u_{in}{}^{m} \varepsilon_{in}{}^{2} \tag{5}$$

In a matrix, J_i is written as

$$J_i = \boldsymbol{\varepsilon}_{\boldsymbol{i}}{}^{T} \mathbf{V}_{\boldsymbol{i}}{}^{m} \boldsymbol{\varepsilon}_{\boldsymbol{i}} \tag{6}$$

Substituting Equation (1) into Equation (6), we obtain:

$$\begin{aligned} J_i &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{\boldsymbol{i}})^{T} \mathbf{V}_{\boldsymbol{i}}{}^{m} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{\boldsymbol{i}}) \\ &= \mathbf{y}^{T} \mathbf{V}_{\boldsymbol{i}}{}^{m} \mathbf{y} - \mathbf{2}\boldsymbol{\beta}_{\boldsymbol{i}}{}^{T} \mathbf{X}^{T} \mathbf{V}_{\boldsymbol{i}}{}^{m} \mathbf{y} + \boldsymbol{\beta}_{\boldsymbol{i}}{}^{T} \mathbf{X}^{T} \mathbf{V}_{\boldsymbol{i}}{}^{m} \mathbf{X} \boldsymbol{\beta}_{\boldsymbol{i}} \end{aligned} \tag{7}$$

To obtain $\widehat{\boldsymbol{\beta}}_\mathrm{i}$, Equation (7) is differentiated with respect to $\widehat{\boldsymbol{\beta}}_\mathrm{i}$ and set equal to zero.

$$\frac{\partial J}{\partial \widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}} = \frac{\partial J_i}{\partial \widehat{\boldsymbol{\beta}}_\mathbf{i}} = -2\mathbf{X}^{T} \mathbf{V}_{\boldsymbol{i}}{}^{m} \mathbf{y} + 2\mathbf{X}^{T} \mathbf{V}_{\boldsymbol{i}}{}^{m} \mathbf{X} \widehat{\boldsymbol{\beta}}_{\boldsymbol{i}} = \mathbf{0} \tag{8}$$

$$\mathbf{X}^{T} \mathbf{V}_{\boldsymbol{i}}{}^{m} \mathbf{y} = \mathbf{X}^{T} \mathbf{V}_{\boldsymbol{i}}{}^{m} \mathbf{X} \widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$$

$$\widehat{\boldsymbol{\beta}}_i = (\mathbf{X}^{T} \mathbf{V}_i{}^{m} \mathbf{X})^{-1} (\mathbf{X}^{T} \mathbf{V}_i{}^{m} \mathbf{y})$$

The second derivative of J is

$$\frac{\partial^2 \boldsymbol{J}}{\partial\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}^{T}} = \frac{\partial^2 \boldsymbol{J}_{\boldsymbol{i}}}{\partial\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}^{T}} = 2\mathbf{X}^{T}\mathbf{V}_{\boldsymbol{i}}^{m}\mathbf{X} \tag{9}$$

The second derivative, $2\mathbf{X}^{\mathrm{T}}\mathbf{V}_{\mathrm{i}}^{\mathrm{m}}\mathbf{X}$, is a positive definite matrix since all diagonal elements in the weight matrix are positive.

The fuzzy weights (*u_{ij})* are updated using a formula obtained through the following process. In Equation (3), for certain *i* and *j* we obtain

$$J_{ij} = u_{ij}{}^{m}\varepsilon_{ij}{}^{2} + \lambda_j(u_{ij} - 1) \tag{10}$$

Deriving Equation (10) with respect to $\mathrm{u_{ij}}$ and setting it equal to zero, we obtain

$$\frac{\partial J}{\partial u_{ij}} = \frac{\partial J_{ij}}{\partial u_{ij}} = m u_{ij}{}^{(m-1)}\varepsilon_{ij}{}^{2} + \lambda_j = 0$$

Thus, $$u_{ij} = \sqrt[m-1]{\frac{-\lambda_j}{m\varepsilon_{ij}{}^{2}}} \tag{11}$$

Since $\sum_{i=1}^{c} \mathrm{u}_{ij} = 1$,

$$\lambda_j = -\left[\left(\sum_{k=1}^{c} \frac{1}{\sqrt[m-1]{m\varepsilon_{kj}{}^{2}}}\right)^{-1}\right]^{1/m-1} \tag{12}$$

Substituting Equation (12) into Equation (11), we obtain:

$$u_{ij} = \sqrt[m-1]{\frac{\left[\left(\sum_{k=1}^{c} \frac{1}{\sqrt[m-1]{m\varepsilon_{kj}{}^{2}}}\right)^{-1}\right]^{1/m-1}}{m\varepsilon_{ij}{}^{2}} \left(\sum_{k=1}^{c}\left[\frac{\varepsilon_{ij}{}^{2}}{\varepsilon_{kj}{}^{2}}\right]^{\frac{1}{m-1}}\right)^{-1}} \tag{13}$$

Statistical tests commonly used in regression analysis cannot be used in FCR. Since the parameter space increases as the number of observations increases, asymptotic properties do not hold, and the distributions of *F* and *t* values are unknown (Cox & Hinkley, 1974). The significance of regression within clusters can be checked with a Monte Carlo significance test via a bootstrap procedure (Wedel & Kistemaker, 1989).

$$\mathrm{SE}^{*}(\hat{\beta}) = \sqrt{\frac{\sum_{\mathrm{b}=1}^{\mathrm{R}}(\hat{\beta}_{\mathrm{b}}^{*} - \bar{\beta}^{*})^{2}}{R-1}} \tag{14}$$

where $$\bar{\beta}^{*} = \sum_{b=1}^{R} \frac{\hat{\beta}_{b}^{*}}{R} \tag{15}$$

R is the number of bootstrap samples, $\widehat{\upbeta}_{\mathrm{b}}^{*}$ is the value of statistic β of the b-th bootstrap sample, and $\bar{\upbeta}^{*}$ is the average of the β values for the bootstrap samples. The t-statistic is

$$t^{*} = \frac{\hat{\beta}_i}{SE^{*}(\hat{\beta}_i)} \tag{16}$$

If the value of $|\mathrm{t}^{*}| > \mathrm{t}_{(\frac{\alpha}{2}, \mathrm{n-k-1})}$ then each predictor variable significantly affects the response variable.

The weighted determination coefficient is used to measure the ability of the predictor variable to explain the response variable (Jajuga, 1986). In the FCR, the weighted determination coefficient is calculated using the formula

$$R_i{}^2 = \frac{\sum_{j=1}^{n} u_{ij}\,(Y_j - \bar{Y})^2 - \sum_{j=1}^{n} u_{ij}\,(Y_j - \hat{Y}_j)^2}{\sum_{j=1}^{n} u_{ij}\,(Y_j - \bar{Y})^2} \tag{17}$$

where $\mathrm{Y_j}$ is the j-th actual response, $\mathrm{\hat{Y}_j}$ is the j-th predicted response, $\overline{\mathrm{Y}}$ is the weighted average of the actual responses, and u_{ij} is the fuzzy membership of the j-th object $(j = 1, 2, \dots, n)$ for the i-th cluster $(i = 1, 2, \dots, c)$.

3. MATERIAL AND METHOD

This study uses a quantitative approach to analyze the relationship between socioeconomic factors such as education, poverty, and unemployment and crime rates in Indonesia. This approach was chosen because it allows researchers to objectively and measurably examine the relationships and statistical patterns between variables. The data used in this study are secondary data obtained from official publications of the Central Statistics Agency (BPS), crime reports from the Indonesian National Police, and education and employment data from relevant agencies for a specific period (e.g., 2020–2022) for each province in Indonesia.

The analytical method used in this study is Fuzzy Clusterwise Regression (FCR), a regression method capable of identifying groups (clusters) in the data that exhibit distinct patterns of regression relationships between independent and dependent variables. FCR was chosen because of its superiority in capturing local heterogeneity and hidden patterns that cannot be uncovered through classical linear regression. In this context, FCR is used to group Indonesian provinces based on their socioeconomic characteristics and estimate different regression models for each group. The dependent variable in this study is the crime rate, while the independent variables include education level, poverty level, and unemployment rate.

The analysis process begins with data pre-processing, such as normalization, multicollinearity testing, and descriptive statistical exploration. Next, a Fuzzy Clusterwise Regression model is estimated by determining the optimal number of clusters using specific criteria such as the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC). Interpretation of the results will focus on differences in relationship patterns between clusters and the contribution of each socioeconomic variable to the crime rate within each group. The results of this study are expected to provide a deeper understanding of the influence of socioeconomic factors on crime and provide data-driven policy input that is more adaptive to regional characteristics.

The following are the stages of analysis: (1) create a descriptive statistical analysis; (2) determine the multiple linear regression parameter model; (3) conduct classical assumption tests and multiple linear regression hypothesis testing; (4) calculate the determination coefficient (5) perform the fuzzy clusterwise regression algorithm; (6) conduct significance testing using the bootstrap approach; (7) calculate the weighted determination coefficient.

Here is the fuzzy clusterwise regression algorithm:

a) At the first iteration $(\mathrm{t} = 0)$
- set $2 \leq \mathrm{c} < \mathrm{n}$, and $\mathrm{m} = 2$;

- generate the initial matrix $\mathbf{U}^{(0)} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_\mathrm{i}, \dots, \mathbf{u}_\mathrm{c})$ where $\mathbf{u}_i{}^T = (u_{\mathrm{i}1}, u_{\mathrm{i}2}, \dots, u_{\mathrm{in}})$

b) Calculate $\widehat{\boldsymbol{\beta}}_\mathrm{i}{}^{(\mathrm{t}+1)} = (\mathbf{X}^\mathrm{T}(\mathbf{V}_\mathrm{i})^\mathrm{t}\mathbf{X})^{-1}(\mathbf{X}^\mathrm{T}(\mathbf{V}_\mathrm{i}{}^\mathrm{m})^\mathrm{t}\mathbf{y})$, with element $\mathbf{V}_\mathrm{i} = \mathrm{diag}(\mathbf{u}_\mathrm{i}{}^\mathrm{t})^\mathrm{m}$, for $\mathrm{i} = 1,2,\dots,\mathrm{c}$.

c) Calculate the residuals $(\boldsymbol{\varepsilon}_\mathrm{i}{}^{(\mathrm{t}+1)})^2 = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_\mathrm{i}{}^{(\mathrm{t}+1)})^2$, for $\mathrm{i} = 1,2,\dots,\mathrm{c}$.

d) Calculate $u_{ij}{}^{(\mathrm{t}+1)} = \left(\sum_{\mathrm{k}=1}^{\mathrm{c}} \left[\frac{(\varepsilon_{\mathrm{ij}}{}^{(\mathrm{t}+1)})^2}{(\varepsilon_{\mathrm{kj}}{}^{(\mathrm{t}+1)})^2}\right]^{\frac{1}{\mathrm{m}-1}}\right)^{-1}$, where $\varepsilon_{kj}{}^{(t+1)} = \frac{1}{(\varepsilon_{ij}{}^{(t+1)})^2}$.

e) Calculate the objective function $J = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij}^{t+1})^m\varepsilon_{ij}^2$ using equation (2).

f) Perform steps (2), (3), (4), and (5) iteratively and stop when the change in the objective function value is less than a predetermined value.

4. RESULTS AND DISCUSSION

Descriptive statistics analysis is a fundamental tool in data analysis, providing a clear and concise overview of a dataset. Summarizing and organizing data, it helps researchers identify patterns, trends, and anomalies. This analysis is essential for understanding the characteristics of a population, making informed decisions, and communicating findings effectively. Table 1 shows the descriptive statistics on the research variables. The provided data shows the descriptive statistics for four variables: the number of crimes, average years of schooling, the number of poor people, and the open unemployment rate. The data is based on a sample of 34 observations. The number of crimes exhibits a wide range, with a minimum of 1220 and a maximum of 51905. The average number of crimes is 10968, with a standard deviation of 12698.21. The average years of schooling are relatively low, with a mean of 9.25 and a standard deviation of 0.82. The number of poor people also shows a significant range, with a mean of 769.40 and a standard deviation of 1082.30. Finally, the open unemployment rate has a mean of 4.99 and a standard deviation of 1.57.

Table 1. Descriptive Statistics

Variable	n	Minimum	Maximum	Mean	Standard Deviation
Number of Crimes (Y)	34	1220	51905	10968.00	12698.21
Average Years of Schooling (X_1)	34	7.31	11.30	9.25	0.82
Number of Poor People (X_2)	34	49.00	4181.00	769.40	1082.30
Open Unemployment Rate (X_3)	34	2.34	8.31	4.99	1.57

Multiple linear regression analysis was conducted to determine the socio-economic factors that influence the number of crimes in Indonesia. The multiple linear regression equation is obtained as follows:

$$Y = -47482.84 + 6034.82X_1 + 10.57\,X_2 - 1099.84X_3$$

Classical assumption testing is carried out later. The normality of residual assumption is met with a value of $\mathrm{D}\ (0{,}17) \leq \mathrm{D}_{(\alpha,\mathrm{n})}(0{,}23)$. In the non-multicollinearity test, the VIF value of the Average Years of Schooling (X_1) is 2.00; the number of poor people (X_2) is 1.55; the open unemployment rate (X_3) is 2.02. These values of VIF are less than 10 means there is no linear relationship between the variables. In the non-autocorrelation test, the value of $d = 2.41$; $d\mathrm{L} = 1.27$; $4 - d\mathrm{L} = 2.73$; $d\mathrm{U} = 1.65$; $4 - d\mathrm{U} = 2.35$ is obtained. The d score is at $4 - d\mathrm{U} < d < 4 - d\mathrm{L}$, so it cannot be concluded. Therefore, a Run Test was conducted and p-value (1.00) > $\alpha\ (0.05)$ was obtained, which means the non-autocorrelation assumption is met. In the homoscedasticity test, the $\mathrm{BP}(6.20\ ) < \mathcal{X}^2_{(\alpha,\mathrm{k}-1)}(7.82)$ means the residual variance is homogeneous.

In the simultaneous test, the F statistic (19.17) is greater than the F table value (2.92). Therefore, the predictor variables affect the response variable simultaneously. In the partial test (t-test), the variable average years of schooling (X_1) obtained a value of $|t|(2.568)$ is greater than $t_{\left(\frac{\alpha}{2},\mathrm{n-k-1}\right)}(2.042)$ and the variable number of poor people (X_2) obtained $|t|(6.762) > t_{\left(\frac{\alpha}{2},\mathrm{n-k-1}\right)}(2.042)$ and p-value $(1.69\text{x}10^{-07}) < \alpha(0.05)$, then these two variables partially have a significant effect on the number of crimes (Y). While the variable open unemployment rate (X_3) obtained a value of $|t|(0.90) < t_{\left(\frac{\alpha}{2},\mathrm{n-k-1}\right)}(2.04)$, we concluded that the variable open unemployment rate (X_3) does not have a significant effect on the number of crimes (*Y*).

The obtained model provides a determination coefficient (R^2) of 0.6572, indicating that 65.72% of the variables of average years of schooling (X_1), number of poor people (X_2), and open unemployment rate (X_3) affect the variability of the number of crimes (*Y*). The remaining 34.28% is influenced by other factors. The R^2 obtained is less than 67% indicating that the model is included in the moderate category. To increase the model performance, the FCR is applied. By using FCR, we will avoid the problem of minimum sample size in each cluster as a problem that arises when using the Spath algorithm (Spath, 1979).

Table 2 presents the average value of the weighted determination coefficient for different numbers of clusters. The weighted determination coefficient is a measure of the goodness of fit of a clustering model. As the number of clusters increases from 2 to 7, the average value of the weighted determination coefficient generally improves, indicating a better fit between the model and the data. The most significant improvement occurs when the number of clusters is 6, where the coefficient is 0.998. Beyond 6 clusters, the improvement becomes marginal, suggesting that adding more clusters might not substantially enhance the model's performance. Unfortunately, when using the number of clusters 5, 6, and 7, a singularity problem arises in the bootstrapping process to estimate the standard error of $\widehat{\boldsymbol{\beta}}$. Therefore, it is finally concluded that the optimal number of clusters in this paper is 4.

Table 2. The Average of the Weighted Determination Coefficient

Number of Clusters (c)	Average Value of Weighted Determination Coefficient
2	0.901
3	0.981
4	0.982
5	0.981
6	0.998
7	0.997

Figure 1 shows the result of grouping all provinces in Indonesia based on a regression model that links the number of crimes with socio-economic factors. Each cluster is characterized by different regression model coefficients. Based on the largest fuzzy weight, the provinces in cluster 1 include West Sumatra, South Sumatra, Jambi, Central Java, Banten, West Nusa Tenggara, East Nusa Tenggara, and North Sulawesi. The regression model in cluster 1 is

$$Y = -118599.857 + 15262.433X_1 + 10.695X_2 - 3636.954X_3$$

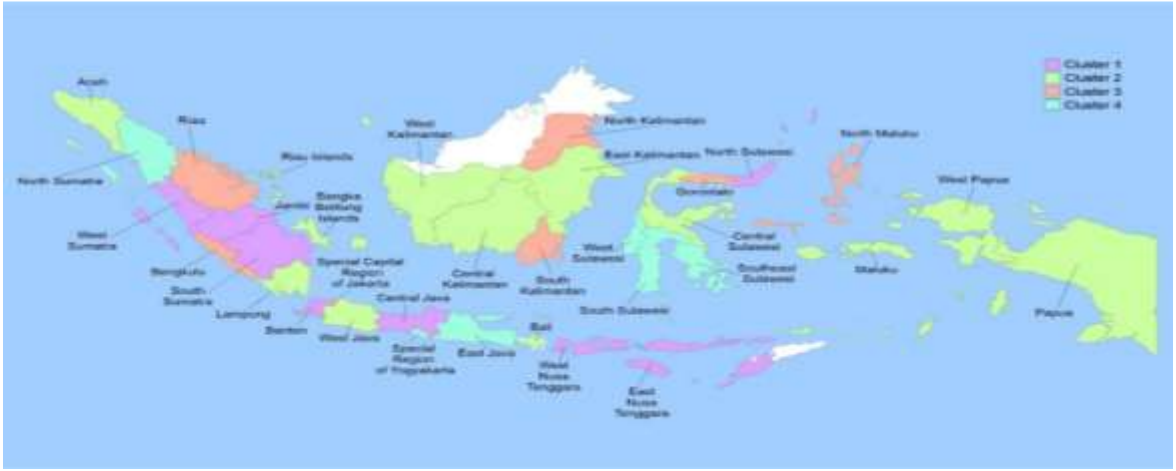

Figure 1. Results of Clustering

The coefficient of average length of schooling (X_1) is 15262.433. This explains that every 1-year increase in the average length of schooling (X_1), will increase the number of crimes (Y) by 15262.433. Next, the coefficient of the number of poor people (X_2) is 10.695. This shows that every 1000 increase in the number of poor people (X_2), will increase the number of crimes (Y) by 10.695. Furthermore, the coefficient of the open unemployment rate (X_3) is -3636.954. This describes that every 1 percent increase in the open unemployment rate (X_3), will reduce the number of crimes (Y) by 3636.954. The weighted determination coefficient in cluster 1 is 0.9720. This means that 97.20% of the variability in the average length of schooling (X_1), the number of poor people (X_2), and the open unemployment rate (X_3) explains the variability in the number of crimes (Y) and the remaining 2.80% is explained by other factors.

According to the largest fuzzy weight, 13 provinces are members of cluster 2. The provinces are Aceh, Lampung, Bangka Belitung Islands, Riau Islands, West Java, Bali, West Kalimantan, Central Kalimantan, East Kalimantan, Central Sulawesi, Maluku, West Papua, Papua. The regression model in cluster 2 is

$$Y = -2703.030 + 893.769X_1 + 7.267X_2 - 624.606X_3$$

The coefficient of the average length of schooling (X_1) is 893.769 which indicates that every 1-year increase in the average length of schooling (X_1) will increase the number of crimes (Y) by 893.769. Then the coefficient of the number of poor people (X_2) is 7.267. This shows that every 1000 increase in the number of poor people (X_2) will increase the number of crimes (Y) by 7.267. Furthermore, the coefficient of the open unemployment rate (X_3) is -624.606 which indicates that every 1 percent increase in the open unemployment rate (X_3) will decrease the number of crimes (Y) by 624.606. The weighted determination coefficient in cluster 1 is 0.979. This means that 97.9% of the variability in the average length of schooling (X_1), the number of poor people (X_2), and the open unemployment rate (X_3) explains the variability in the number of crimes (Y) and the remaining 2.1 % is explained by other factors.

Based on the largest fuzzy weight, cluster 3 consists of the provinces of Riau, Bengkulu, DKI Jakarta, South Kalimantan, North Kalimantan, Gorontalo, and North Maluku. The regression model in cluster 3 is

$$Y = -54315.781 + 3649.508X_1 + 24.401X_2 + 4628.370X_3$$

The coefficient of average length of schooling (X_1) is 3649.508. This can be interpreted that every 1-year increase in the average length of schooling (X_1) will increase the number of crimes (Y) by 3649.508. Furthermore, the coefficient of the number of poor people (X_2) is

24.401, which means that every 1000 increase in the number of poor people (X_2) will increase the number of crimes (Y) by 24.401. Then the coefficient of the open unemployment rate (X_3) is 4628.370. This informs that every 1 percent increase in the open unemployment rate (X_3) will increase the number of crimes (Y) by 4628.370. The weighted determination coefficient in cluster 1 is 0.980. This means that 98% of the variability in the average length of schooling (X_1), the number of poor people (X_2), and the open unemployment rate (X_3) explains the variability in the number of crimes (Y) and the remaining 2% is explained by other factors.

According to the largest fuzzy weight, cluster 4 consists of 6 provinces. These provinces include North Sumatra, DI Yogyakarta, East Java, South Sulawesi, Southeast Sulawesi, and West Sulawesi

$$Y = 60609.83 - 11053.79X_1 + 0.636X_2 + 15045.52X_3$$

The coefficient of the average length of schooling (X_1) is -11053.79 which shows that every 1-year increase in the average length of schooling (X_1) will reduce the number of crimes (Y) by 11053.79. Then the coefficient of the number of poor people (X2) is 0.636. This explains that every 1000 increase in the number of poor people (X_2) will increase the value of the number of crimes (Y) by 0.636. Furthermore, the coefficient of the open unemployment rate (X_3) is 15045.52 which means that every 1 percent increase in the open unemployment rate (X_3) will increase the number of crimes (Y) by 15045.52. The weighted determination coefficient in cluster 1 is 0.995. This means that 99.5% of the variability in the average length of schooling (X_1), the number of poor people (X_2), and the open unemployment rate (X_3) explains the variability in the number of crimes (Y) and the remaining 0.5% is explained by other factors.

Table 3. Partial Test on Parameters of Each Cluster

Variable	Parameter estimates	Bootstrap Std. Error	t^*	$t_{(0,025,30)}$	Sig.
Cluster 1					
Average Years of Schooling (X_1)	15262.433	238.968	63.868	2.042	Yes
Number of Poor People (X_2)	10.695	0.088	122.019	2.042	Yes
Open Unemployment Rate (X_3)	-3636.954	77.623	-46.854	2.042	Yes
Cluster 2					
Average Years of Schooling (X_1)	893.769	60.203	14.846	2.042	Yes
Number of Poor People (X_2)	7.267	0.097	74.859	2.042	Yes
Open Unemployment Rate (X_3)	-624.606	0.283	-14.784	2.042	Yes
Cluster 3					
Average Years of Schooling (X_1)	3649.508	112.403	32.468	2.042	Yes
Number of Poor People (X_2)	24.401	0.283	86.140	2.042	Yes
Open Unemployment Rate (X_3)	4628.370	66.714	69.377	2.042	Yes
Cluster 4					
Average Years of Schooling (X_1)	-11053.79	950.083	-11.635	2.042	Yes
Number of Poor People (X_2)	0.636	1.487	0.428	2.042	No
Open Unemployment Rate (X_3)	15045.52	671.829	22.395	2.042	Yes

Table 3 presents the results of individual tests for regression coefficients in each of the four clusters. These tests assess the statistical significance of each independent variable (Average Years of Schooling, Number of Poor People, and Open Unemployment Rate) in predicting the dependent variable. In general, the results indicate that most of the independent variables are statistically significant predictors of the dependent variable in all four clusters. This is evidenced by the "Yes" values in the "Sig." column for the majority of the coefficients. For example, in Cluster 1, all three independent variables have significant coefficients, suggesting that they are all important predictors of the dependent variable. However, there is one exception in Cluster 4. The coefficient for the "Number of Poor People" in this cluster is not significant, as indicated by the "No" in the "Sig." column. This implies that the number of poor people is not a statistically significant predictor of the dependent variable in Cluster 4.

5. CONCLUSION

FCR can increase the coefficient of determination from 65.72% in the multiple linear regression model to more than 90% in FCR. FCR is not only to improve model performance but also to find the different relationships between variables in each cluster. There are several differences in the results of the significance test in classical regression and FCR. In the classical regression model, the open unemployment rate does not affect on the number of crimes. However, after clustering using the Fuzzy Clusterwise Regression method with the number of clusters of 4, variables that partially have a significant effect on the number of crimes in cluster 1, cluster 2, and cluster 3 are average years of schooling, the number of poor people, and the open unemployment rate. Meanwhile, in cluster 4, the variable of the number of poor people partially has no significant effect on the number of crimes.

ACKNOWLEDGMENT

This research was supported by Research Grant “Riset Pengembangan dan Penerapan (RPP)” with contract number 973-10/UN7.D2/PP/IX/2024 by Institute for Research and Community Services, Diponegoro University.

REFERENCES

Arsono, Y. D., & Atmanti, H. D. (2014). Pengaruh Variabel Pendidikan, Pengangguran, Rasio Gini, Usia, dan Jumlah Polisi Perkapita Terhadap Angka Kejahatan Properti di Provinsi Jawa Tengah Tahun 2010-2012. *Undergraduate Thesis*, Universitas Diponegoro.

Bertsekas, D. P. (1982). *Constrained Optimization and Lagrange Multiplier Methods*. New York: Academic Press. https://doi.org/10.1016/B978-0-12-093480-5.50007-6

BPS. (2022). Statistik Kriminal 2022. *Badan Pusat Statistik*, *023*, 30–80. https://doi.org/4401002

BPS. (2023). *Statistik Kriminal 2023*. Badan Pusat Statistik.

Breetzke, G. D., & Pearson, A. L. (2015). Socially Disorganized Yet Safe: Understanding Resilience to Crime in Neighborhoods in New Zealand. *Journal of Criminal Justice*, *43*(6), 444–452. https://doi.org/10.1016/j.jcrimjus.2015.09.001

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical Statistics*. New York: Chapman and Hall. https://doi.org/https://doi.org/10.1201/b14832

Desarbo, W. S., & Cron, W. L. (1988). A Maximum Likelihood Methodology for Clusterwise Linear Regression. In *Journal of Classification*, *5*, 249-282

Goulas, E., & Zervoyianni, A. (2015). Economic Growth and Crime: Is There An Asymmetric Relationship? *Economic Modelling*, *49*(September), 286–295. https://doi.org/10.1016/j.econmod.2015.04.014

Grover, C. (2012). *Crime and Inequality*. UK: Routledge.

Jajuga, K. (1986). Linear Fuzzy Regression. *Fuzzy Sets and Systems*, *20*(3), 343–353. https://doi.org/10.1016/S0165-0114(86)90045-X

Kartono, K. (2009). *Patologi Sosial Jilid 1* (2nd ed.). Rajawali Pers.

Klawoon, F., & Hoppner, F. (2003). What is Fuzzy About Fuzzy Clustering? Understanding and Improving the Concept of the Fuzzifier. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2810*(May), 5. https://doi.org/10.1007/978-3-540-45231-7

Lochner, L. (2020). Education and Crime. *The Economics of Education: A Comprehensive Overview*, 109–117. https://doi.org/10.1016/B978-0-12-815391-8.00009-4

Lochner, L., & Moretti, E. (2004). The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports. *The American Economic Review*, *94*(1), 155–189.

Melick, M. D. (2003). The Relationship Between Crime and Unemployment. *The Park Place Economist*, *11*(1), 30–36.

Mohammed, H., & Mohamed, W. A. W. (2015). Reducing Recidivism Rates through Vocational Education and Training. *Procedia - Social and Behavioral Sciences*, *204*(November 2014), 272–276. https://doi.org/10.1016/j.sbspro.2015.08.151

O'Sullivan, A. (2019). *Urban Economics* (9th ed.). McGraw-Hill.

Pare, P. P., & Felson, R. (2014). Income Inequality, Poverty and Crime Across Nations. *British Journal of Sociology*, *65*(3), 434–458. https://doi.org/10.1111/1468-4446.12083

Setiadi, E. (2000). Reformasi Hukum Pidana, untuk Mengantisipasi Perkembangan Kejahatan di Bidang Ekonomi (Economic Crimes. *Mimbar Jurnal Sosial Dan Pembangunan*, *16*(3), 205–214. https://doi.org/10.1080/10611991.2016.1251223

Soekanto, S., Liklikuwata, H., & Kusumah, M. W. (1986). *Kriminologi Suatu Pengantar*. Ghalia Indonesia.

Spath, H. (1979). Clusterwise Linear Regression. *Computing*, *22*, 367–373.

Wahyuni P., D. (2010). Mencermati Perilaku Kekerasan dan Paradigma Sosial. *Unisia*, *0*(61 SE-Articles), 339–349. https://doi.org/10.20885/unisia.vol29.iss61.art9

Wedel, M., & Kistemaker, C. (1989). Consumer Benefit Segmentation using Clusterwise Linear Regression. *International Journal of Research in Marketing*, *6*(1), 45–59. https://doi.org/10.1016/0167-8116(89)90046-3

Wedel, M., & Steenkamp, J. B. E. M. (1989). A Fuzzy Clusterwise Regression Approach to Benefit Segmentation. *International Journal of Research in Marketing*, *6*(4), 241–258. https://doi.org/10.1016/0167-8116(89)90052-9

Wu, K. L. (2012). Analysis of Parameter Selections for Fuzzy C-Means. *Pattern Recognition*, *45*(1), 407–415. https://doi.org/10.1016/j.patcog.2011.07.012