DOI: https://doi.org/10.34123/jurnalasks.v18i1.972

POLITEKNIK STATISTIKA STIS

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 18 No. 1, June 2026
e- ISSN: 2615-1367, p-ISSN:2086-4132

Available online at https://jurnal.stis.ac.id/

JASKS Jurnal Aplikasi Statistika & Komputasi Statistik

Mapping and Modeling Crime Factors in North Sumatra Using GWGPR

Eva Kosasih[1*], Ni Luh Putu Suciptawati[2], Luh Putu Ida Harini[3]

[1,2,3]Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Udayana, Bali, Indonesia
**Corresponding Author: E-mail address: kosasih.2208541054@student.unud.ac.id*

ARTICLE INFO

Article history:
Received 4 May, 2026
Revised 15 June, 2026
Accepted 25 June, 2026
Published 30 June, 2026

Keywords:
Crime Cases; Geographically Weighted Generalized Poisson Regression; North Sumatra; Overdispersion; Spatial Heterogeneity; Spatial Modeling

Abstract

Introduction/Main Objectives: Crime remains a significant social issue influenced by socio-economic factors and exhibiting spatial variation, particularly in North Sumatra Province, which recorded the highest number of criminal cases in Indonesia in 2024. This study aims to identify significant factors affecting crime and examine the spatial variation of their effects across districts/cities in North Sumatra. **Background Problems:** Global regression models often fail to capture crime patterns due to overdispersion and spatial heterogeneity, leading to inconsistent relationships across regions. **Novelty:** This study employs Geographically Weighted Generalized Poisson Regression (GWGPR), which simultaneously addresses overdispersion and spatial heterogeneity, providing a more robust localized analysis than global models. **Research Methods:** Using secondary data from 33 districts/cities in North Sumatra, the variables include population density, open unemployment rate, mean years of schooling, and Gini ratio. The analysis involves Poisson regression, dispersion testing, Generalized Poisson Regression, spatial heterogeneity testing, and GWGPR. **Finding/Results:** The significant factors affecting crime are the open unemployment rate, mean years of schooling, and population density, while the Gini ratio is not significant. **Limitation:** This study is limited by the use of data covering only the year 2024 and a limited set of socio-economic variables, which may not fully capture all factors associated with crime.

1. Introduction

Crime remains a major social problem in both developed and developing countries because it threatens public security, social stability, and community welfare [1]. Various types of crime, such as theft, robbery, fraud, assault, and property-related offenses, continue to occur in many regions. In Indonesia, North Sumatra is one of the provinces with a relatively high crime rate. According to data from the National Criminal Information Center of the Indonesian National Police, North Sumatra recorded 53,897 criminal cases in 2024, the highest number among all provinces in Indonesia [2]. This condition highlights the need for effective policy interventions to address crime in North Sumatra.

Crime is influenced not only by individual behavior but also by socio-economic conditions. From a theoretical perspective, crime can be explained through several socio-economic theories. Social

Disorganization Theory suggests that areas characterized by high population concentration tend to experience weaker informal social control and greater opportunities for criminal activities because increased social interaction and urban concentration may reduce the effectiveness of community supervision [3]. Consistent with this theory, previous studies have reported that population density has a significant effect on crime rates, as densely populated areas increase social interaction and competition for resources [4]. Furthermore, Economic Theory of Crime argues that individuals make rational decisions by comparing the expected benefits and costs of legal and illegal activities. Under unfavorable economic conditions, such as unemployment and limited access to formal employment opportunities, individuals may face stronger incentives to engage in criminal behavior [5]. Consistent with this argument, empirical evidence suggests that open unemployment contributes to increased criminal behavior due to economic pressure and limited job opportunities [6]. Consequently, the open unemployment rate is considered an important explanatory variable.

Human Capital Theory also emphasizes that education improves skills, legal awareness, and employment opportunities, thereby reducing the likelihood of criminal involvement [7]. Consistent with this theory, education level, represented by mean years of schooling, is expected to influence crime rates. As reported in [8], mean years of schooling is associated with lower crime rates, potentially due to improved legal awareness and broader economic opportunities. In addition, Relative Deprivation Theory proposes that income inequality may generate perceptions of social injustice and economic exclusion that encourage criminal behavior [9]. Consistent with this perspective, previous studies have shown that income inequality, measured by the Gini ratio, can lead to social dissatisfaction and increase incentives for criminal behavior [10], [11]. Based on these theoretical arguments and previous findings, population density, open unemployment rate, mean years of schooling, and income inequality are hypothesized to influence crime in this study.

Given the theoretical and empirical evidence discussed above, the selection of population density, open unemployment rate, mean years of schooling, and the Gini ratio is particularly relevant to North Sumatra. These variables represent key demographic, labor market, educational, and inequality dimensions that have been widely associated with crime in previous studies. Since socio-economic conditions differ across regions, the relationships between crime and socio-economic factors may exhibit geographical variation and spatial dependence [12]. Consequently, the effects of these socio-economic factors are unlikely to be uniform across locations, and both the magnitude and direction of their relationships with crime may vary spatially [12], [13]. Therefore, a local modeling approach is more appropriate than a global model for capturing the heterogeneous relationships between crime and the factors influencing it across districts and cities [14]. However, spatial heterogeneity is not the only characteristic that should be considered when modeling crime data.

In addition to spatial heterogeneity, the number of criminal cases also exhibits count-data characteristics that should be considered when selecting an appropriate statistical model. Crime data, specifically the number of criminal cases, are classified as count data. Poisson regression is commonly used for modeling count data. However, it assumes equidispersion, where the variance equals the mean. In practice, crime data often exhibit overdispersion or underdispersion, leading to inefficient estimation and biased statistical inference [15]. Generalized Poisson Regression (GPR) can address this limitation because it includes an additional dispersion parameter [16]. However, GPR is a global model that assumes identical relationships across all regions. This assumption may be unrealistic in practice, as districts and cities in North Sumatra have different socio-economic characteristics, indicating spatial heterogeneity.

To simultaneously accommodate dispersion and spatial heterogeneity, this study applies Geographically Weighted Generalized Poisson Regression (GWGPR). This method enables local parameter estimation for each district and city while handling dispersion in count data. Previous studies have shown that GWGPR performs better than Poisson regression and GPR models in spatial count-data applications [17], [18]. However, no study has specifically applied GWGPR to district-level crime data in North Sumatra. Therefore, this study aims to model criminal cases, identify significant factors influencing crime, and map their spatial variation across districts and cities in North Sumatra in 2024 using GWGPR.

2. Material and Methods

2.1. Poisson Regression

Poisson regression is commonly used for modeling count data, where the response variable is discrete and assumed to follow a Poisson distribution. A discrete random variable Y follows a Poisson distribution with parameter μ if it has the following probability mass function [19]:

$$P(Y = y \mid \mu) = \frac{e^{-\mu}\mu^{y}}{y!}, y = 0,1,2,\ldots \quad (1)$$

Suppose that $Y_i \sim \text{Poisson}(\mu_i)$, for $i = 1,2,\ldots,n$. The Poisson regression model uses the natural logarithm as the link function to relate the expected value of Y_i to explanatory variables x_i. The model can be written as:

$$\ln\big(E(Y_i)\big) = \mathbf{x}_i^T\boldsymbol{\beta} \quad (2)$$

or equivalently,

$$E(Y_i) = \mu_i = \exp\big(\mathbf{x}_i^T\boldsymbol{\beta}\big) \quad (3)$$

where $\mathbf{x}_i = [1 \;\; x_{i1} \;\; x_{i2} \;\; \ldots \;\; x_{ik}]^T$ is the predictor vector and $\boldsymbol{\beta} = [\beta_0 \;\; \beta_1 \;\; \beta_2 \;\; \ldots \;\; \beta_k]^T$ is the parameter vector.

Parameter estimation in Poisson regression is performed using Maximum Likelihood Estimation (MLE), with the likelihood function given by:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \quad (4)$$

where $\mu_i = \mathrm{e}^{\mathbf{x}_i^T\boldsymbol{\beta}}$.

2.2. Multicollinearity

Multicollinearity refers to the presence of high correlation among predictor variables, where one predictor variable can be linearly explained by other predictors. The detection of multicollinearity can be carried out using the Variance Inflation Factor (VIF), which is defined as follows [20]:

$$VIF_j = \frac{1}{1 - R_j^2}, j = 1,2,\ldots,k \quad (5)$$

where R_j^2 is the coefficient of determination obtained by regressing the j-th predictor on the remaining predictors. Variables with VIF values greater than 5 indicate multicollinearity.

2.3. Dispersion Test

Poisson regression assumes equidispersion, meaning that the mean equals the variance of the response variable. Violation of this assumption may lead to overdispersion or underdispersion, which can result in inefficient parameter estimation and biased standard errors. To detect dispersion, the deviance and Pearson chi-square statistics are compared with their respective degrees of freedom [21]. The dispersion statistics are defined as follows:

$$\phi_1 = \frac{D}{df}, \qquad D = 2\sum_{i=1}^{n}\left(y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right) \quad (6)$$

$$\phi_2 = \frac{\chi^2}{df}, \qquad \chi^2 = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \quad (7)$$

where $df = n - k - 1$, n is the number of observations, and k is the number of predictor variables. The variable y_i denotes the observed value, while $\hat{\mu}_i$ represents the predicted value. The dispersion condition is assessed based on the value of ϕ. A value of $\phi < 1$ indicates underdispersion, $\phi = 1$ indicates equidispersion, and $\phi > 1$ indicates overdispersion.

2.4. Generalized Poisson Regression (GPR)

Generalized Poisson Regression (GPR) is an extension of the Poisson regression model used to analyze count data that do not satisfy the equidispersion assumption. This model introduces an additional dispersion parameter (ϕ) to accommodate both overdispersion and underdispersion [22]. The probability mass function of the generalized Poisson distribution is given as follows:

$$P(Y_i = y_i) = \left(\frac{\mu_i}{1+\phi\mu_i}\right)^{yi} \frac{(1+\phi y_i)^{y_i-1}}{y_i!} \exp\left[-\frac{\mu_i(1+\phi y_i)}{1+\phi\mu_i}\right] \tag{8}$$

where $y_i = 0,1,2,\ldots$, $\mu_i = \mathrm{e}^{\mathbf{x}_i^T\boldsymbol{\beta}}$, and ϕ is the dispersion parameter. The mean and variance are given by:

$$E(Y_i) = \mu_i,\ Var(Y_i) = \mu_i(1+\phi\mu_i)^2$$

Parameter estimation in the GPR model is performed using Maximum Likelihood Estimation (MLE). The likelihood function of the GPR model is given by:

$$L(\phi, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\mu_i}{1+\phi\mu_i}\right)^{y_i} \frac{(1+\phi y_i)^{y_i-1}}{y_i!} \exp\left(-\frac{\mu_i(1+\phi y_i)}{1+\phi\mu_i}\right) \tag{9}$$

Simultaneous testing of model parameters is performed using the Maximum Likelihood Ratio Test (MLRT). The hypotheses are defined as follows [23]:

$H_0 :\ \beta_1 = \beta_2 = \cdots = \beta_k = 0$

$H_1 :$ at least one $\beta_j \neq 0; j = 1,2,\ldots,k$

The test statistic is given by:

$$D(\hat{\beta}) = -2\ln\left(\frac{L(\hat{\omega})}{L(\hat{\Omega})}\right) = 2\left[\ln L(\hat{\Omega}) - \ln L(\hat{\omega})\right] \tag{10}$$

where $L(\hat{\omega})$ is the likelihood of the restricted model without predictors, and $L(\hat{\Omega})$ is the likelihood of the full model. The decision rule is to reject H_0 if $D(\hat{\beta}) > \chi^2_{(\propto,k)}$.

Partial parameter testing is performed using the Wald test [24]. The hypotheses are:

$H_0 :\ \beta_j = 0, j = 1,2,\ldots,k$

$H_1 :\ \beta_j \neq 0, j = 1,2,\ldots,k$

The test statistic is given by:

$$W = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}\right)^2 \tag{11}$$

where $\hat{\beta}_j$ is the estimated parameter and $SE(\hat{\beta}_j)$ is its standard error. The decision rule is to reject H_0 if $W > \chi^2_{(\propto;1)}$.

2.5. Spatial Heterogeneity

Spatial heterogeneity is a condition in which the relationships between variables in a regression model vary across observation locations. The presence of spatial heterogeneity is assessed using the Breusch–Pagan (BP) test, with the following hypotheses:

H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2$

H_1: at least one $\sigma_i^2 \neq \sigma^2,\ i = 1,2,..,n$

The Breusch–Pagan (BP) test statistic is defined as:

$$BP = \left(\frac{1}{2}\right)\mathbf{f}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{f} \tag{12}$$

The elements of vector **f** are defined as $f_i = \left(\frac{\varepsilon_i^2}{\sigma^2} - 1\right)$ where $\varepsilon_i = y_i - \hat{y}_i$, and Z is an $n \times (2k+1)$ matrix consisting of a constant term and interaction terms between predictor variables x_{ij} and spatial coordinates (u_i, v_i), namely $x_{ij}u_i$and $x_{ij}v_i$. The decision rule is to reject H_0 if $BP > \chi^2_{(\alpha,df)}$ where α is the significance level, $df = 2k$ and k is the number of predictors variables.

2.6. Spatial Weight Matrix

Spatial variability across locations is represented through the construction of a spatial weight matrix W. Each element of the matrix is defined based on the Euclidean distance between locations. In GWGPR, the spatial weight matrix is constructed using kernel functions that assign weights to neighboring observations according to their spatial proximity to a target location. Observations located closer to the target location receive larger weights, whereas observations located farther away receive lower weights. The weighting process is controlled by a bandwidth parameter, which determines the spatial extent of neighboring observations involved in local parameter estimation. The choice of bandwidth is important because it affects the balance between local and global information and influences the accuracy of model estimation [14].

Kernel functions can generally be classified into fixed kernels and adaptive kernels. Fixed kernels use a constant bandwidth for all locations, meaning that neighboring observations are weighted based on a fixed geographical distance. Consequently, the number of neighboring observations included in each local model may vary across locations. In contrast, adaptive kernels use a variable bandwidth that adjusts according to the spatial distribution of observations. As a result, the number of neighboring observations remains relatively constant, while the geographical distance represented by the bandwidth may vary across locations. Therefore, the choice of kernel function depends on the spatial characteristics of the data [14]. In the GWGPR model, several commonly used kernel functions are as follows:

a. Fixed Gaussian

$$w_{ij} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{h}\right)^2\right) \tag{13}$$

b. Adaptive Gaussian

$$w_{ij} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{h_j}\right)^2\right) \tag{14}$$

c. Fixed Bisquare

$$w_{ij} = \begin{cases} \left(1 - \left(\frac{d_{ij}}{h}\right)^2\right)^2, & d_{ij} \le h \\ 0, & d_{ij} > h \end{cases} \tag{15}$$

d. Adaptive Bisquare

$$w_{ij} = \begin{cases} \left(1 - \left(\frac{d_{ij}}{h_j}\right)^2\right)^2, & d_{ij} \le h_j \\ 0, & d_{ij} > h_j \end{cases} \tag{16}$$

The variable d_{ij} represents the Euclidean distance between location i and j, defined as :

$$d_{ij} = \sqrt{\left(u_i - u_j\right)^2 + \left(v_i - v_j\right)^2} \tag{17}$$

where (u_i, v_i) denotes the geographical coordinates (latitude and longitude) of location i. The parameter h represents the bandwidth, which determines the spatial range of neighboring observations influencing local parameter estimation. For adaptive kernels, the bandwidth varies across locations and is denoted by h_j. The selection of an optimal bandwidth is crucial because it affects model accuracy as well as the bias–variance trade-off in local estimation [14].

The optimal bandwidth can be determined using the Cross-Validation (CV) method, which is defined as follows [14]:

$$CV = \sum_{i=1}^{n} \left(y_i - \hat{y}_{\neq i}(h)\right)^2 \quad (18)$$

where $\hat{y}_{\neq i}(h)$ is the estimated value of y_i obtained by excluding the observation at location (u_i, v_i), and n is the number of observations. The optimal bandwidth is selected by minimizing the CV value, where a smaller CV value indicates better predictive performance[14].

2.7. Geographically Weighted Generalized Poisson Regression (GWGPR)

Geographically Weighted Generalized Poisson Regression (GWGPR) model is an extension of the Generalized Poisson Regression model that allows parameter estimates to vary across locations. In this model, the regression coefficients are estimated locally for each observation point based on its geographical coordinates (u_i, v_i). The GWGPR model for the i-th location using the log-link function is expressed as:

$$\eta_i = g(\mu_i) = \ln(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}(u_i, v_i)$$
$$\mu_i = g^{-1}(\eta_i) = \exp\left(\mathbf{x}_i^T \boldsymbol{\beta}(u_i, v_i)\right) \quad (19)$$

Parameter estimation in the GWGPR model is carried out using Maximum Likelihood Estimation (MLE) method. The likelihood function is given by [10]:

$$L(\boldsymbol{\beta}(u_i, v_i), \phi) = \prod_{i=1}^{n} \left(\frac{\mu_i}{1+\phi\mu_i}\right)^{y_i} \frac{(1+\phi y_i)^{y_i-1}}{y_i!} \exp\left[\frac{-\mu_i(1+\phi y_i)}{1+\phi\mu_i}\right] \quad (20)$$

where $\mu_i = \exp\left(\mathbf{x}_i^T \boldsymbol{\beta}(u_i, v_i)\right)$

Simultaneous testing of model parameters is conducted using the Maximum Likelihood Ratio Test (MLRT), with hypotheses [23]:

H_0 : $\beta_1(u_i, v_i) = \beta_2(u_i, v_i) = \cdots = \beta_k(u_i, v_i) = 0$

H_1 : at least one $\beta_j(u_i, v_i) \neq 0$; $j = 1,2,\ldots,k$

The test statistic is defined as:

$$D\left(\hat{\beta}(u_i, v_i)\right) = -2\ln\left(\frac{L(\widehat{\omega})}{L(\widehat{\Omega})}\right) = 2\left[\ln L(\widehat{\Omega}) - \ln L(\widehat{\omega})\right] \quad (21)$$

where $L(\widehat{\omega})$ is the likelihood of the restricted model without predictors, and $L(\widehat{\Omega})$ is the likelihood of the full model. The decision rule is to reject H_0 if $D(\hat{\beta}) > \chi^2_{(\propto,k)}$.

Partial testing of parameters is conducted using the Wald test [24]. The hypotheses are:

H_0 : $\beta_j(u_i, v_i) = 0$, $j = 1,2,\ldots,k$; $i = 1,2,\ldots,n$

H_1 : $\beta_j(u_i, v_i) \neq 0$, $j = 1,2,\ldots,k$; $i = 1,2,\ldots,n$

The test statistic is given by:

$$W = \left(\frac{\hat{\beta}_j(u_i, v_i)}{SE\left(\hat{\beta}_j(u_i, v_i)\right)}\right)^2 \quad (22)$$

where $\hat{\beta}_j(u_i, v_i)$ is the estimated regression coefficient at location (u_i, v_i), and $SE\left(\hat{\beta}_j(u_i, v_i)\right)$ is its standard error. The decision rule is to reject H_0 if $W > \chi^2_{(\propto;1)}$.

2.8. Data Source

This study uses secondary data from 33 districts/cities in North Sumatra Province, Indonesia, consisting of 25 regencies and 8 cities. Crime data for 2024 were obtained from the National Criminal Information Center of the Indonesian National Police (Pusiknas Polri), while predictor variables were obtained from Statistics Indonesia (BPS) of North Sumatra Province. The variables used in this study are summarized in Table 1.

Table 1. Research variables

Variable	Description	Unit
Y	Number of criminal cases	Cases
X_1	Population density	Persons/km^2
X_2	Open unemployment rate	Percent
X_3	Mean years of schooling	Years
X_4	Gini ratio	Index

2.9. Analysis Steps

The data analysis is performed using R and GeoDa software. The analytical procedure is carried out as follows:

1. Descriptive Analysis
   Descriptive statistics are computed to describe the characteristics of criminal cases in North Sumatra Province in 2024, along with the factors suspected of influencing them.
2. Multicollinearity Detection
   Multicollinearity among predictor variables is assessed using the Variance Inflation Factor (VIF) based on equation (5). A predictor variable is considered to exhibit multicollinearity if $VIF > 5$. Variable selection is performed by removing predictors with high VIF values or strong correlations until a set of independent variables is obtained.
3. Poisson regression Modeling
   (a) Estimating the Poisson regression model parameters using the Maximum Likelihood Estimation (MLE) method based on equation (4).
   (b) Conducting dispersion testing using equations (6) and (7) to identify overdispersion or underdispersion. If no dispersion issues are detected, the Poisson regression model is considered adequate; otherwise, the analysis proceeds to Generalized Poisson Regression (GPR).
4. Generalized Poisson Regression Modeling
   (a) Estimating the GPR model parameters using the MLE method based on equation (9).
   (b) Testing parameter significance simultaneously using the Maximum Likelihood Ratio Test (MLRT) (10) and partially using the Wald test (11).
5. Spatial Heterogeneity Testing
   Spatial heterogeneity is tested using the Breusch–Pagan test based on equation (12). If spatial heterogeneity is detected, the analysis proceeds to Geographically Weighted Generalized Poisson Regression (GWGPR) modeling; otherwise, the GPR model is considered sufficient.
6. Geographically Weighted Generalized Poisson Regression Modeling
   (a) Calculating the Euclidean distance between observation locations based on geographic coordinates (latitude and longitude) using equation (17).
   (b) Determining the optimum bandwidth for the kernel weighting function using the Cross-Validation (CV) method (18).
   (c) Constructing the weighting matrix with the selected kernel function using equations (13), (14), (15), and (16).
   (d) Estimating the GWGPR model parameters using the MLE method based on equation (20).
   (e) Testing parameter significance simultaneously using the MLRT (21) and partially using the Wald test (22).

3. Results and Discussion

3.1. Descriptive Analysis

Descriptive statistics of the number of criminal cases and the factors presumed to influence them in North Sumatra Province in 2024 are presented in Table 2, while the spatial distribution of criminal cases is illustrated in Figure 1.

Table 2. Descriptive statistics of research variables

Variable	Min	Max	Mean	Std Dev
Y	13	16,791	1,488.394	2,872.636
X_1	41.16	8,902.16	1,159.908	2,221.838
X_2	0.89	8.13	4.358	2.372
X_3	6.4	11.82	9.503	1.322
X_4	0.206	0.356	0.259	0.040

Based on Table 2, the average number of criminal cases across districts/cities in North Sumatra in 2024 is 1,488.39, with a large standard deviation (2,872.64) and a wide range from 13 to 16,791 cases, indicating substantial variation across regions. This pattern is evident in Figure 1, which shows that Medan City records the highest number of cases, while Gunungsitoli City has the lowest. In particular, criminal cases are concentrated in the administrative and economic center of the province, especially in Medan City, as reflected by the darker orange color, while variations in color gradation suggest that crime is not evenly distributed across districts/cities in North Sumatra.

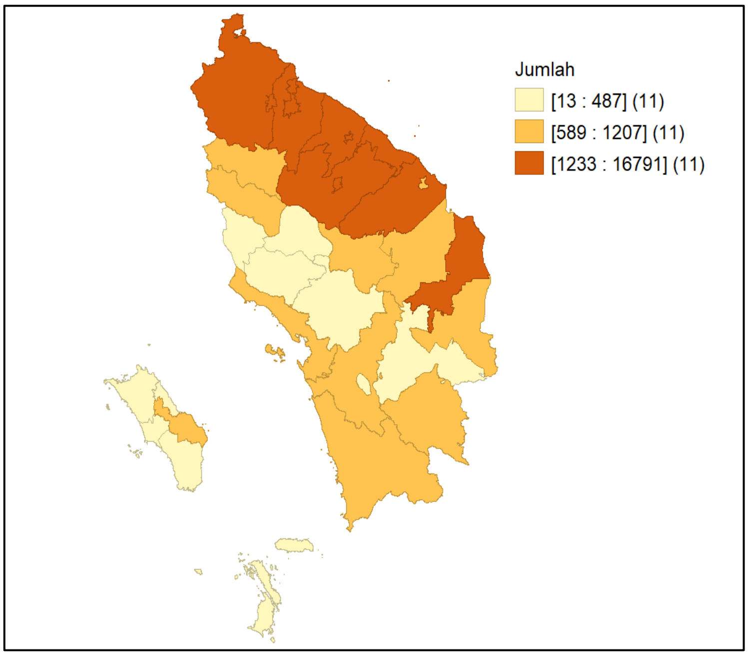

Figure 1. Spatial distribution of the number of criminal cases in North Sumatra

3.2. Multicollinearity Testing

Multicollinearity among predictor variables is assessed using the Variance Inflation Factor (VIF), with a threshold of $VIF < 5$. The results are presented in Table 3.

Table 3. VIF values of predictor variables

Predictor Variable (X)	X_1	X_2	X_3	X_4
VIF Values	4.18	2.46	1.90	3.26

The VIF values in Table 3 indicate that all predictor variables have values below 5. This suggests that there is no multicollinearity among the predictors, and all variables can be included in the subsequent modeling analysis.

3.3. Poisson Regression Modeling

An initial analysis was conducted using Poisson regression, since crime data is count data. Based on the estimated parameters, the Poisson regression model is expressed as follows:

$$\hat{\mu} = \exp(4.286 + 0.0001193X_1 + 0.2577X_2 + 0.007367X_3 + 4.772X_4) \quad (24)$$

The Poisson regression model assumes equidispersion, meaning that the mean equals the variance. To assess this assumption, dispersion testing was performed by comparing the deviance and Pearson chi-square statistics to their respective degrees of freedom. The results are presented in Table 4.

Table 4. Dispersion test result for Poisson Regression Model

Method	Statistic	df	Ratio (Statistic/df)
Deviance	29,126.71	28	1,040.24
Pearson Chi-Squares	26,274.17	28	938.36

As shown in Table 4, both the deviance ratio and the Pearson chi-square ratio are substantially greater than 1. This indicates the presence of overdispersion, where the variance exceeds the mean. Consequently, the Poisson model is not appropriate for this data, and an alternative model that can accommodate overdispersion, such as Generalized Poisson Regression (GPR), is required.

3.4. Generalized Poisson Regression Modeling

Following the detection of overdispersion in the Poisson model, the analysis proceeds with Generalized Poisson Regression (GPR). Using the Maximum Likelihood Estimation (MLE) method, the model parameters were estimated, resulting in the following equation:

$$\hat{\mu} = \exp(5.880 + 0.00001348X_1 + 0.1778X_2 + 0.2753X_3 - 8.171X_4) \quad (25)$$

To evaluate the reliability of the model, parameter significance was assessed both simultaneously and partially. The simultaneous test was conducted using the Maximum Likelihood Ratio Test (MLRT) with the following hypotheses:

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
H_1 : at least one $\beta_j \neq 0; j = 1,2,3,4$

The test produced a deviance value of 21.0627, which exceeds the critical chi-square value of $\chi^2_{(0,05;4)} = 9.4877$. Therefore, the null hypothesis (H_0) is rejected, indicating that at least one predictor variable significantly affects the response variable.

Furthermore, the significance of individual predictors was examined using the Wald test. The hypotheses are defined as:

$H_0 : \beta_j = 0, j = 1,2,3,4$
$H_1 : \beta_j \neq 0, j = 1,2,3,4$

The results of the Wald test are presented in Table 5.

Table 5. Wald Test results for the GPR Model

Parameter	Wald Value (W)	$\chi^2_{(0,05;1)}$	Conclusion
β_1	0.031	3.841	Not Significant
β_2	7.272	3.841	Significant
β_3	4.718	3.841	Significant
β_4	2.750	3.841	Not Significant

As shown in Table 5, the open unemployment rate (X_2) and mean years of schooling (X_3) have Wald statistics greater than the critical value of 3.841. Therefore, the null hypothesis (H_0) is rejected for these variables, indicating that they have a statistically significant effect on the number of criminal cases. In contrast, population density (X_1) and the Gini ratio (X_4) have Wald statistics less than the critical value. Thus, the null hypothesis fails to be rejected for these variables, implying that they do not have a statistically significant effect in the model.

3.5. Spatial Heterogeneity Testing

Spatial heterogeneity testing was conducted to determine whether the relationship between predictor variables and the number of criminal cases is consistent across locations or varies spatially. Spatial heterogeneity was assessed using the Breusch–Pagan (BP) test with the following hypotheses:

$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_{33}^2 = \sigma^2$

H_1: at least one $\sigma_i^2 \neq \sigma^2,\ i = 1,2,..,33$

The analysis produced a BP statistic value of 25.7462, which exceeds the critical chi-square value of $\chi^2_{(0,05;8)} = 15.507$. Therefore, the null hypothesis (H_0) is rejected, indicating the presence of spatial heterogeneity in the data. This result implies that the relationships between predictor variables and criminal cases vary across districts/cities in North Sumatra. Consequently, the GPR model is considered inadequate, and the Geographically Weighted Generalized Poisson Regression (GWGPR) model is employed to account for spatial variation and to provide location-specific parameter estimates.

3.6. Geographically Weighted Generalized Poisson Regression Modeling

The significance of spatial heterogeneity, as identified by the Breusch–Pagan test, indicates that the relationship between predictor variables and the number of criminal cases varies across locations. Therefore, the GPR model is insufficient because it assumes that the effects of explanatory variables are constant across all locations. To accommodate both spatial heterogeneity and overdispersion in count data, the Geographically Weighted Generalized Poisson Regression (GWGPR) model is employed.

The GWGPR modeling process begins by calculating the Euclidean distance between observation locations using geographical coordinates (longitude and latitude). These distances are subsequently used to construct a spatial weight matrix that reflects the influence of neighboring observations on local parameter estimation. The weighting structure is determined through a kernel function and an associated bandwidth parameter. The optimal bandwidth is selected using the Cross-Validation (CV) method. The results of bandwidth selection for each kernel function are presented in Table 6.

Table 6. Optimum bandwidth selection

Kernel Function	Cross Validation (CV)	Bandwidth	Base Bandwidth
Adaptive Gaussian	388,093,455	0.7272668	24 nearest neighbors
Fixed Gaussian	366,850,232	63.57038	63.57 km
Adaptive Bisquare	355,129,647	0.7272577	24 nearest neighbors
Fixed Bisquare	354,705,234	224.5921	224.59 km

Based on Table 6, the adaptive kernel functions (Adaptive Gaussian and Adaptive Bisquare) produce optimal bandwidths corresponding to approximately 24 nearest neighbors. This represents about 72.7% of the 33 districts/cities included in the analysis, indicating that each local model incorporates information from a substantial proportion of the available observations. Consequently, local parameter estimation is based on a relatively large number of neighboring observations. In contrast, the fixed kernel functions (Fixed Gaussian and Fixed Bisquare) use constant bandwidths of 63.57 km and 224.59 km, respectively, indicating that the weighting scheme is determined based on a fixed geographical distance rather than a fixed number of neighboring observations. Among the four candidate kernel functions, the Fixed Bisquare kernel produces the smallest Cross Validation (CV) value of 354,705,234. Since the CV criterion aims to minimize prediction error, the kernel function with the lowest CV value is considered to provide the most appropriate spatial weighting structure for the data. Therefore, the Fixed Bisquare kernel was selected for constructing the spatial weight matrix and estimating the GWGPR model, with an optimal bandwidth of 224.59 km.

Using the Maximum Likelihood Estimation (MLE) method, local parameters are estimated for each district/city. As an example, the GWGPR model for Langkat Regency is expressed as follows:

$$\hat{\mu}_{\text{Langkat}} = \exp(12.9372 + 0.0004X_1 + 0.2387X_2 - 0.4173X_3 - 11.9047X_4) \qquad (26)$$

To validate the model, significance testing is conducted. The simultaneous test using the Maximum Likelihood Ratio Test (MLRT) is defined with the following hypotheses:

$H_0 : \beta_1(u_i, v_i) = \beta_2(u_i, v_i) = \beta_3(u_i, v_i) = \beta_4(u_i, v_i) = 0$
H_1 : at least one $\beta_j(u_i, v_i) \neq 0;\ j = 1,2,3,4$

The test produced a deviance value of 203.5904, which exceeds the critical chi-square value of $\chi^2_{(0,05;4)} = 9.4877$. Therefore, the null hypothesis (H_0) is rejected, indicating that at least one predictor variable significantly affects the response variable. Furthermore, the significance of individual predictors is evaluated using the Wald test. The hypotheses are defined as:

$H_0 : \beta_j(u_i, v_i) = 0,\ j = 1,2,3,4; \quad i = 1,2, \dots ,33$
$H_1 : \beta_j(u_i, v_i) \neq 0,\ j = 1,2,3,4; \quad i = 1,2, \dots ,33$

The Wald test results for Langkat Regency are presented in Table 7.

Table 7. Wald Test results for the GWGPR in Langkat Regency model

Parameter	Wald Value (W)	$\chi^2_{(0,05;1)}$	Conclusion
β_1	19.7110	3.841	Significant
β_2	6.5652	3.841	Significant
β_3	4.5411	3.841	Significant
β_4	2.8829	3.841	Not Significant

As shown in Table 7, the Wald statistics for population density (X_1), open unemployment rate (X_2), and mean years of schooling (X_3) are greater than the critical value of 3.841. Therefore, the null hypothesis (H_0) is rejected for these variables, indicating that they have a statistically significant effect on the number of criminal cases. In contrast, the Gini ratio (X_4) has a Wald statistic less than the critical value. Thus, the null hypothesis fails to be rejected, implying that this variable does not have a statistically significant effect in the model. Therefore, the local GWGPR model for Langkat Regency can be expressed as follows:

$$\hat{\mu}_{\text{Langkat}} = \exp(12.9372 + 0.0004X_1 + 0.2387X_2 - 0.4173X_3) \qquad (27)$$

Based on the estimated model (27), population density (X_1) has a positive effect on the number of criminal cases, with a coefficient of 0.0004, corresponding to $\exp(0.0004) = 1.0004$. This indicates that a one-unit increase in population density (persons/km²), holding other variables constant, increases the expected number of criminal cases by approximately 1.0004 times. Similarly, the open unemployment rate (X_2) shows a positive effect, with a coefficient of 0.2387 ($\exp(0.2387) = 1.2696$), implying that a 1% increase in unemployment increases the expected number of criminal cases by approximately 1.2696 times. In contrast, mean years of schooling (X_3) has a negative effect, with a coefficient of −0.4173 ($\exp(-0.4173) = 0.6588$), indicating that an increase of one year in schooling reduces the expected number of criminal cases to approximately 0.6588 times. These findings are consistent with previous studies, which suggest that higher population density and unemployment increase crime due to intensified social interaction and economic pressure, while higher education reduces crime through improved legal awareness and broader economic opportunities [4], [6], [8]. A summary of significant variables across districts and cities in North Sumatra is presented in Table 8.

As shown in Table 8, the open unemployment rate (X_2) is the most consistently significant variable across regions, indicating that economic pressure plays an important role in driving criminal activity. This finding is consistent with previous research [6], which reported that increases in unemployment tend to be accompanied by increases in criminal activity due to the economic pressures faced by individuals. In contrast, the Gini ratio (X_4) is not significant in any district/city. One possible explanation is that the variation in income inequality across districts/cities is relatively low, as indicated by the low standard deviation reported in Table 2. Consequently, the observed differences in income inequality may not be sufficiently large to explain spatial variations in crime rates.

Table 8. Summary of significant predictors in 33 districts/cities

Significant Variable	Districts/Cities
X_1, X_2	Medan
X_1, X_2, X_3	Langkat
X_2	Asahan, Batu Bara, Binjai, Dairi, Deli Serdang, Humbang Hasundutan, Karo, Labuhan Batu Utara, Nias Selatan, Pakpak Bharat, Pematangsiantar, Samosir, Serdang Bedagai, Simalungun, Tanjung Balai, Tapanuli Utara, Tebing Tinggi, Toba
X_3	Nias Barat, Nias Utara
No Significant Variables	Nias, Mandailing Natal, Tapanuli Selatan, Tapanuli Tengah, Labuhanbatu, Padang Lawas Utara, Padang Lawas, Labuhan Batu Selatan, Sibolga, Padangsidempuan, dan Gunungsitoli

However, this finding does not necessarily imply that income inequality has no relationship with crime in North Sumatra. The insignificant effect may also reflect the presence of other socio-economic factors, such as population density and unemployment, which exhibit stronger associations with criminal activity in the study area. In addition, the Gini ratio is an aggregate indicator that may not fully capture localized socio-economic disparities relevant to criminal behavior.

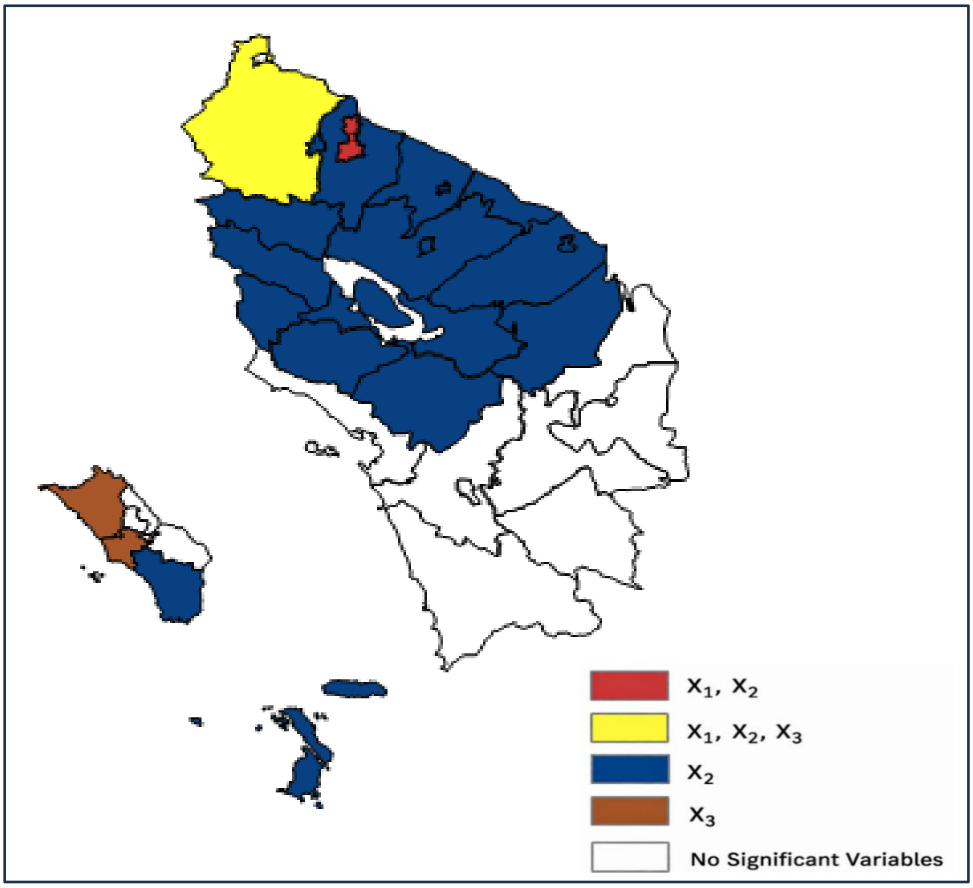

Figure 2. Spatial distribution of significant predictor variables based on the GWGPR model in North Sumatra

This result is consistent with previous findings reported in [25], which indicated that the relationship between income inequality and crime was not always statistically significant. However, it differs from the findings reported in [10], where income inequality was identified as a significant determinant of crime in Indonesia. These differences may be attributable to variations in study area, period of observation, data characteristics, and modeling approaches. Therefore, the relationship between income inequality and crime may vary across regions and depend on local socio-economic conditions.

Additionally, several districts/cities exhibit no significant predictors. This suggests that variations in crime within these areas may be associated with other factors not included in the present model, such as poverty, urbanization, law enforcement effectiveness, demographic structure, or other socio-economic characteristics. However, the absence of significant predictors in some districts/cities does not imply that the GWGPR model is invalid. Instead, it reflects the spatial heterogeneity of crime determinants, indicating that the selected explanatory variables do not have statistically significant local effects in every location. This finding highlights that crime patterns may be influenced by different factors across regions and supports the use of GWGPR, which allows the relationships between crime and explanatory variables to vary across locations. The spatial distribution of significant predictors is illustrated in Figure 2.

4 Conclusion

The results of the GWGPR model indicate that the effects of predictor variables vary across locations, confirming the presence of spatial heterogeneity. Among the predictors, the open unemployment rate (X_2) is the most dominant factor, being statistically significant in 22 districts/cities, followed by mean years of schooling (X_3) in 3 districts/cities and population density (X_1) in 2 districts/cities. These findings demonstrate that the significant factors affecting crime vary across districts and cities in North Sumatra.

The spatial variation identified by the GWGPR model has important policy implications. In districts/cities where unemployment is the only significant factor, policies should focus on employment creation, vocational training, and entrepreneurship programs. In Medan, where both population density and unemployment are significant, crime prevention strategies should combine labor market interventions with urban management and public security measures in densely populated areas. In Langkat, where population density, unemployment, and education are all significant, integrated policies addressing education, employment, and community-based crime prevention are required. Meanwhile, in Nias Barat and Nias Utara, where education is the only significant predictor, efforts to improve educational access and reduce school dropout rates may contribute to crime reduction.

Several districts/cities do not exhibit significant effects for any of the variables included in the model. This suggests that crime in these areas may be influenced by other factors not considered in the present study, such as poverty, urbanization, demographic characteristics, or law enforcement effectiveness. Therefore, for future research, it is recommended to explore additional variables that may influence crime patterns and consider models that integrate both global and local effects, such as Mixed Geographically Weighted Generalized Poisson Regression (MGWGPR).

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

Derived data supporting the findings of this study are available from the corresponding author on request.

Credit Authorship

Eva Kosasih: Conceptualization, Methodology, Data Collection, Data Analysis, Writing – Original Draft. **Ni Luh Putu Suciptawati:** Research Advisor, Writing – Review. **Luh Putu Ida Harini:** Research Advisor, Writing – Review.

References

[1] S. Bayu and Y. Hudi, "Dampak Kriminalitas Terhadap Kualitas Hidup Masyarakat Urban," Media Hukum Indonesia, vol. 2, no. 6, pp. 308–312, 2025.
[2] Pusiknas Polri, Statistik Kriminal Periode 01 Jan 2024 - 31 Des 2024. Jakarta, Indonesia: Pusiknas Polri, 2025.
[3] C. R. Shaw and H. D. McKay, Juvenile Delinquency and Urban Areas. Chicago, IL, USA: University of Chicago Press, 1942.
[4] J. A. Parenja, M. A.-Z. Salsabila, and Fatmawati, "Dampak Kepadatan Penduduk di Pulau Jawa Terhadap Kesejahteraan dan Kriminalitas," Pediaqu: Jurnal Pendidikan Sosial dan Humaniora, vol. 4, no. 2, pp. 4328–4337, 2025.
[5] G. S. Becker, "Crime and Punishment: An Economic Approach," Journal of Political Economy, vol. 76, no. 2, pp. 169–217, 1968.
[6] K. N. Azmi, S. P. Azzahra, V. K. Dewi, and Y. W. Pertiwi, "Analisis Pengangguran Terhadap Tindakan Kriminalitas di Kota Bekasi," Observasi: Jurnal Publikasi Ilmu Psikologi, vol. 2, no. 3, pp. 223–234, 2024.
[7] T. W. Schultz, "Investment in Human Capital," American Economic Review, vol. 51, no. 1, pp. 1–17, 1961.
[8] D. R. A. Syarifuddin, Pengaruh Rata-Rata Lama Sekolah dan Upah Minimum Terhadap Kriminalitas Melalui Kemiskinan di Indonesia, Bachelor's thesis, Universitas Islam Negeri Alauddin Makassar, Makassar, Indonesia, 2023
[9] R. A. Stouffer, Society and Education. Boston, MA, USA: Allyn & Bacon, 1955.
[10] N. A. Simangunsong, "Pengaruh Ketimpangan Pendapatan Terhadap Kejahatan di Indonesia," Indonesian Journal of Economics and Strategic Management, vol. 2, no. 2, pp. 1612–1620, 2024.
[11] L. Sugiharti, R. Purwono, M. A. Esquivias, and H. Rohmawati, "The Nexus between Crime Rates, Poverty, and Income Inequality: A Case Study of Indonesia," Economies, vol. 11, no. 2, pp. 1–15, 2023.
[12] P. A. Widyastaman and D. Hartono, "Geographic Distribution of Economic Inequality and Crime in Indonesia: Exploratory Spatial Data Analysis and Spatial Econometrics Approach," Applied Spatial Analysis and Policy, vol. 17, no. 2, pp. 547–571, 2024.
[13] L. Sugiharti, M. A. Esquivias, M. S. Shaari, L. Agustin, and H. Rohmawati, "Criminality and Income Inequality in Indonesia," Social Sciences, vol. 11, no. 3, p. 142, 2022.
[14] A. S. Fotheringham, C. Brunsdon, and M. Charlton, Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Chichester, UK: Wiley, 2002.
[15] S. Q. Aina, M. Fauziyah, and W. P. Nurmayanti, "Comparison of Generalized Poisson Regression and Negative Binomial Regression Models Based on Akaike Information Criterion Values," Statistika, vol. 25, no. 1, pp. 86–101, 2025.
[16] M. J. Badriawan and S. Melaniani, "Aplikasi generalized Poisson regression untuk memodelkan faktor yang mempengaruhi jumlah kasus baru difteri di Provinsi Jawa Timur," Media Gizi Kesmas, vol. 12, no. 2, pp. 860–869, 2023.
[17] S. W. Tyas and L. A. Puspitasari, "Geographically Weighted Generalized Poisson Regression Model with the Best Kernel Function in the Case of the Number of Postpartum Maternal Mortality in East Java," MethodsX, vol. 10, p. 102002, 2023.
[18] A. Uswah, J. Rizal, and Y. Fauzi, "Comparison of Geographically Weighted Generalized Poisson Regression (GWGPR) and Geographically Weighted Negative Binomial Regression

(GWNBR) Methods in Determining Factors Affecting Tuberculosis Cases in Indonesia," J. Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika, vol. 18, no. 1, pp. 851–865, 2025.

[19] A. C. Cameron and P. K. Trivedi, Regression Analysis of Count Data, 2nd ed. Cambridge, UK: Cambridge University Press, 2013.

[20] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to Linear Regression Analysis, 5th ed. Hoboken, NJ, USA: Wiley, 2012.

[21] I. Z. Safira, "Penerapan Metode Generalized Poisson Regression pada Jumlah Kriminalitas di Provinsi Jawa Barat," B.S. thesis, Universitas Islam Negeri Sunan Ampel Surabaya, Surabaya, Indonesia, 2023.

[22] C. C. I. A. Njudang, R. D. Guntur, E. B. Sinu, and J. R. Pahnael, "Perbandingan Model Generalized Poisson Regression dan Negative Binomial Regression dalam Mengatasi Overdispersi pada Kasus Pneumonia di Provinsi NTT Tahun 2023," MATH UNESA: Jurnal Ilmiah Matematika, vol. 13, no. 3, pp. 88–96, 2025.

[23] P. McCullagh and J. A. Nelder, Generalized Linear Models, 2nd ed. London, UK: Chapman and Hall, 1989.

[24] A. Agresti, An Introduction to Categorical Data Analysis, 3rd ed. Hoboken, NJ, USA: Wiley, 2019.

[25] D. T. Anozi and B. Novianda, "Socio-Economic and Property Crime Rate in Indonesia," Economics Development Analysis Journal, vol. 12, no. 3, pp. 305–318, 2023.