BAREKENG: Journal of Mathematics and Its Applications
December 2024 Volume 18 Issue 4 Page 2181–2190
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss4pp2181-2190

MODELING LONGITUDINAL FLOOD DATA IN WEST SUMATRA USING THE GENERALIZED ESTIMATING EQUATION (GEE) APPROACH

Alfi Nur Nitasari[1], Andini Sa'idah[2], Nurin Faizun[3], Kezia Eunike Darmawan[4], Marfa Audilla Fitri[5], Nur Chamidah[6*]

[1,2,3,4,5]*Statistics Study Program, Mathematics Department, Faculty of Science and Technology, Universitas Airlangga*

[6]*Mathematics Department, Faculty of Science and Technology, Universitas Airlangga*
Jl. Dr. Ir. H. Soekarno, Mulyorejo, Surabaya, East Java 60115, Indonesia

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

Article History:

Received: 1st January 2024
Revised: 11th May 2024
Accepted: 7th July 2024
Published: 14th October 2024

Keywords:

Flood;
West Sumatera;
Poisson Regression;
Negative Binomial Regression;
Generalized Estimating Equation (GEE).

ABSTRACT

Flooding is one of the many natural disasters that often hit Indonesia. In July 2023, three areas in West Sumatra experienced floods and landslides which caused damages and even 2 missing victims. Since November 16th, 2023, 8 hamlets in Meranti Village, Landak District, West Sumatra have been inundated by floods which affected 359 *families and many public facilities. This research uses data from West Sumatra Province Central Statistics Agency. The data used is 2014, 2018 and 2021. The response variable used is the number of villages/sub-districts experiencing natural disasters according to district/city (y). The predictor variables used are regional topography* (x_1)*, the number of water channels such as rivers, reservoirs, etc.* (x_2)*, the number of fields cleared through burning* (x_3)*, the number of villages/sub-districts in C excavation area* (x_4)*, and the number of dumpsters* (x_5)*. This research uses Negative Binomial Regression with the Generalized Estimating Equation (GEE) approach. In the Poisson regression test, the QIC value based on Independent Working Correlation Structure (WCS) is* 263.365 *with deviance value of* 263.02*, degree of freedom of* 57*, and dispersion score of 4,6144. Because the dispersion value is greater than 1, it can be concluded that there is overdispersion. Because there is more than one overdispersion, it is overcome by using negative binomial. The results of parameter estimation using negative binomial regression based on Independent WCS showed that only one variable was significant, which is the number of fields cleared through burning* (x_3) *with deviance value of* 34.61*, degrees of freedom of* 57 *and a QIC of* 39.51*. Negative Binomial regression model that was formed is* $y = exp\,(1.383329 + 0.0234557x_3)$*. From the two regression models used, namely Poisson and negative binomial, it was found that the negative binomial regression model was the best model because it had the lowest QIC value of* 39.51.

How to cite this article:

A. N. Nitasari, A. Sa'idah, N. Faizun, K. E. Darmawan, M. A. Fitri and N. Chamidah., "MODELING LONGITUDINAL FLOOD DATA IN WEST SUMATRA USING THE GENERALIZED ESTIMATING EQUATION (GEE) APPROACH," *BAREKENG: J. Math. & App.*, vol. 18, iss. 4, pp. 2181-2190, December, 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

Indonesia is one of the countries located in the ring of fire, which is the joint of three tectonic plates, the Indo-Australian Plate, the Eurasian Plate and the Pacific Plate. Hence, Indonesia is a country prone to natural disasters. A natural disaster is a natural occurrence that has a major impact on the human population [1].

One of the disasters that often hit Indonesia is floods. Flooding can take the form of stagnant, puddling water on agricultural land, city centers, or residential areas that should be kept dry. Floods also occur when the debit or volume of water through a drainage channel or river exceeds its drainage capacity. Increasing water debit and volume should not be a big problem as long as it does not cause harm, cause fatalities, injuries, or cause other problems in daily life [2]. Unfortunately, floods often not only cause losses such as forcing residents to evacuate, but can also damage public facilities and block road access. In July 2023, three areas in West Sumatra, namely Agam Regency, Padang Pariaman and Padang City, were hit by floods and landslides. This disaster caused various damages to facilities and infrastructure, blocked road access, and even resulted in two missing victims [3]. Other than that, since November 16th, 2023, it was reported that Meranti Village, Landak District, West Sumatera has been soaked [4]. This flood affected 359 families and many public facilities. The flood phenomenon in West Sumatra is an event that occurs year after year. West Sumatra has diverse topography, ranging from lowlands to mountains, with high rainfall. This area is also located along the Sumatra fault line, contributing to a high risk of natural disasters such as floods and landslides. Flood modeling in this region needs to consider the complex spatial and temporal variability

With so much damage, loss and casualties caused by floods, it is important to get to the root of the problem to help prepare for and prevent future floods. There are three comprehensive steps that can help prevent flooding in flood risk management, namely identification, assessment and risk control. The identification step can analyze vulnerable areas based on factors that may affect them such as water conditions, surrounding land conditions, and others. Then, the assessment step is used to determine the amount of risk that will be faced by considering factors such as area, population, and others. Finally, the risk control step can be carried out flood mitigation efforts such as building embankments, planting trees around waters to slow down the flow of water, making terraces to reduce water runoff, and others [5]. In this research, the first step is to analyze the factors that cause flooding with several existing journal references. In the previous research on flood frequency in East Java on 2011-2013 using Generalized Estimating Equation (GEE) method [6] said that regional topography, the number of water channels (rivers, reservoirs, etc.), the number of fields cleared through burning, the number of villages/sub-districts in C excavation area, and the number of dumpsters can be causes of flood.

The data used in this research is in count form, so Poisson Regression can be used. If there is overdispersion, Negative Binomial regression can be used as a standard model for count data and the model is included in the non-linear regression model with the variables used being the number of villages or sub-districts that experienced flooding, the number of villages or districts based on regional topography, the number of villages or sub-districts based on water channels, the number of villages and districts doing fields cleared through burning, the number of villages or sub-districts based on class C excavation locations, and the number of villages or sub-districts based on the availability of dumpsters in the year 2014, 2018, and 2021.

Longitudinal data is commonly used in both observational and experimental studies. In longitudinal studies, individuals in the study are followed up over a certain period of time for each individual, so that data is collected at several points in time [7]. Longitudinal data is characterized by the fact that repeated observations within the same subject tend to be correlated, so that models for the analysis of longitudinal data must take into account the relationships between periodic observations within the same subject.

Generalized Estimating Equation (GEE), introduced by Liang and Zeger in 1986, is a development of the Generalized Linear Model (GLM) method which can be used to estimate model parameters that are autocorrelated and not normally distributed [8]. This method is semi-parametric because the estimating equations are derived without full specification of the combined distribution of the observation subjects. The GEE method requires selecting the appropriate correlation structure to describe the correlation. This correlation structure selection uses Quasi-Likelihood Under the Independence Information Criterion (QIC). Previous research shows GEE handles non-independent data well, as seen in hydrology, and demonstrates the long-term impacts of climate and environmental variables, making it a robust choice for longitudinal data with significant correlations.

The GEE approach effectively analyzes longitudinal flood data in West Sumatra by accounting for time-based correlations and changes in flood patterns. It integrates covariates like rainfall, land use, and climate change, offering a more comprehensive analysis than simpler models. With all existing considerations, it is decided that Negative Binomial Regression using GEE on longitudinal data on the frequency of floods in West Sumatra in 2014, 2018, and 2021 will be used in the research entitled "Modeling Longitudinal Flood Data in West Sumatra using the Generalized Estimating Approach Equation (GEE)".

2. RESEARCH METHODS

2.1 Generalized Estimating Equation (GEE)

Generalized Estimating Equation (GEE) is another regression approach for a longitudinal study that is used to analyze longitudinal data with correlated data [6]. The GEE method is a development of GLM which can be used to estimate model parameters based on data that contains autocorrelation and data that is not normally distributed. Longitudinal data is often characterized by correlations between repeated observations on the same subject, where GEE explicitly accounts for these correlations through the selected working matrix and can be used for various types of response variables, both continuous and categorical allowing us to analyze different types of data [9]. GEE provides consistent estimators for the regression parameters, even if the assumed correlation structure is incorrect. This makes GEE a very useful method in practice [10]. GEE has linear predictor specifications as follows:

$$\eta_{ij} = x_i\beta \tag{1}$$

With link function of $g(\eta_{ij}) = \eta_{ij}$. The types and choises of link function in GEE is the same as in GLM. The variance of response variable y is determined by:

$$V(y_{ij}) = \phi v(\mu_{ij}) \tag{2}$$

With μ_{ij} a variance function and ϕ a known or estimated scale parameter. A specification from GEE that is not found in GLM is a specification for the correlation between two different responses or is often referred to as a working correlation structure $\boldsymbol{R_i}$ sized $n_i \times n_i$ which is dependent of parameter α, therefore this correlation matrix is often called as $\boldsymbol{R_i(\alpha)}$ [11]. GEE addresses the problem of correlation in longitudinal data by using a working matrix to account for the correlation structure between repeated observations on the same subject to illustrate how observations at different times are correlated with each other [9]. GEE provides consistent estimators for regression parameters, even if the assumed correlation structure is not completely correct. GEE provides a wide selection of correlation matrices, allowing the researcher to choose the correlation matrix that best fits the data.. Some forms of correlation matrices that are often used in GEE are as follows:

Table 1. Working Correlation Matrix

Structure	Definition	Number of Parameters
Independent	$R_{u,v} = \begin{cases} 1, \text{if } u = v \\ 0, \text{otherwise} \end{cases}$ $\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$	0
Exchangeable	$R_{u,v} = \begin{cases} 1, \text{ if } u = v \\ \rho, \text{ otherwise} \end{cases}$ $\begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$	1

Structure	Definition	Number of Parameters
	$R_{u,v} = \begin{cases} 1, & \text{if } u = v \\ \rho_{u,v}, & \text{otherwise} \end{cases}$	
Unstructured	$\begin{pmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,k} \\ \rho_{1,2} & 1 & \cdots & \rho_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,k} & \rho_{2,k} & \cdots & 1 \end{pmatrix}$	$\frac{k(k-1)}{2}$

2.2 Negative Binomial Distribution

If over-dispersion occurs in discrete data, using Poisson regression is still ineffective because the standard error value becomes lower than expected. This is because the regression coefficient parameters of Poisson regression are inefficient, even though the regression coefficients remain consistent.

The Negative Binomial Distribution is used to form a Negative Binomial regression model, where Negative Binomial regression does not emphasize the equidispersion assumption required in Poisson regression **[12]**.To form a regression model, mean and variance are obtained in the form of $E(y) = \mu$ and $Var(y) = \mu(1 + \theta\mu)$ using the mass probability function of negative binomial as follows:

$$f(y;\mu,\theta) = \frac{\Gamma\left(y + \frac{1}{\theta}\right)}{\Gamma\left(\frac{1}{\theta}\right) \text{y!}} \left(\frac{1}{1+\theta\mu}\right)^{\frac{1}{\theta}} \left(\frac{\theta\mu}{1+\theta\mu}\right)^{y} \tag{3}$$

The negative binomial distribution will approximate a Poisson distribution if the mean and variance are assumed to be the same, namely $E(y) = Var(y) = \mu$. The distribution function of the exponential family of the negative binomial distribution is as follows:

$$f(y;\mu,\theta) = \exp\left\{ y \ln\left(\frac{\theta\mu}{1+\theta\mu}\right) + \frac{1}{\theta}\ln\left(\frac{1}{1+\theta\mu}\right) + \ln\left(\frac{\Gamma\left(y + \frac{1}{\theta}\right)}{\Gamma\left(\frac{1}{\theta}\right) \text{y!}}\right)\right\} \tag{4}$$

The contribution of predictor variables in the negative binomial regression model is expressed in the form of a linear combination between the parameters (η) with regression parameters to be estimated as follows:

$$\eta_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} \tag{5}$$

$$\eta = X\beta \tag{6}$$

With η a vector of size $(n \times 1)$ from observation, X is a matrix $(n \times c)$ of predictor variables, β is a matrix $(c \times 1)$ of regression coefficients with $c = p + 1$, and $E(y)$ is discrete and has a positive value. Therefore, to transform the value of η_i (a real number) into a pool that corresponds to the range of responses y, a link function $g(.)$ is required, which is:

$$\begin{aligned} g(\mu_i) &= \ln \mu_i, \text{with } \mu_i = \exp\left(X_i{}^T \beta\right) \text{ then} \\ g(\mu_i) &= \ln\left(exp\left(X_i{}^T \beta\right)\right) \\ g(\mu_i) &= X_i{}^T \beta \end{aligned} \tag{7}$$

The maximum likelihood method with the Newton-Raphson procedure is used to estimate the parameters of the negative binomial regression **[13]**. Using this method requires the first and second derivatives of the likelihood function, $\boldsymbol{y_i}$ has a negative binomial distribution probability mass function as follows.

$$\boldsymbol{f(y_i|\mu_i,\theta) = \frac{\Gamma\left(y_i + \frac{1}{\theta}\right)}{\Gamma\left(\frac{1}{\theta}\right)\Gamma(y_i+1)} \left(\frac{1}{1+\theta\mu_i}\right)^{\frac{1}{\theta}} \left(\frac{\theta\mu_i}{1+\theta\mu_i}\right)^{y_i}} \tag{8}$$

2.3 Data Source and Variables

The data used in this research are data on the frequency of flood events, data on rubbish dumps, regional topography data, and data on the number of water channels such as rivers, reservoirs, etc. Apart from that, data on the number of fields cleared by burning and the number of villages/sub-districts located in the C excavation area are also used. The data used is from year 2014, 2018 and 2021 and focuses on the province of West Sumatra. This data is secondary data obtained from the official website of Badan Pusat Statistik Provinsi Sumatera Barat This research has two variables consisting of five predictor variables and one response variable. All research variables are presented in **Table 2** as follows:

Table 2. Research Variables

Variable Type	Variable Name	Data Type
Response (y)	Number of Villages/Subdistricts Experiencing Natural Disasters According to Regency/City	Continuous
Predictor (x_1)	Number of Subdistricts/Villages/Nagari According to Regency/City and Regional Topography [Slope/Peak + Valley + Expanse/Plain]	Continuous
Predictor (x_2)	Number of Villages / Subdistricts by District/City and Existence of [Rivers, Irrigation Channels, Lakes/Reservoirs/Dams, Springs]	Continuous
Predictor (x_3)	Number of Villages/Subdistricts by fields cleared through burning	Continuous
Predictor (x_4)	Number of villages/Subdistricts by District/City and existence of C excavation area.	Continuous
Predictor (x_5)	Number of Villages/Kelurahan by Regency/City and Availability of Dumpsters	Continuous

2.4 Analysis Steps

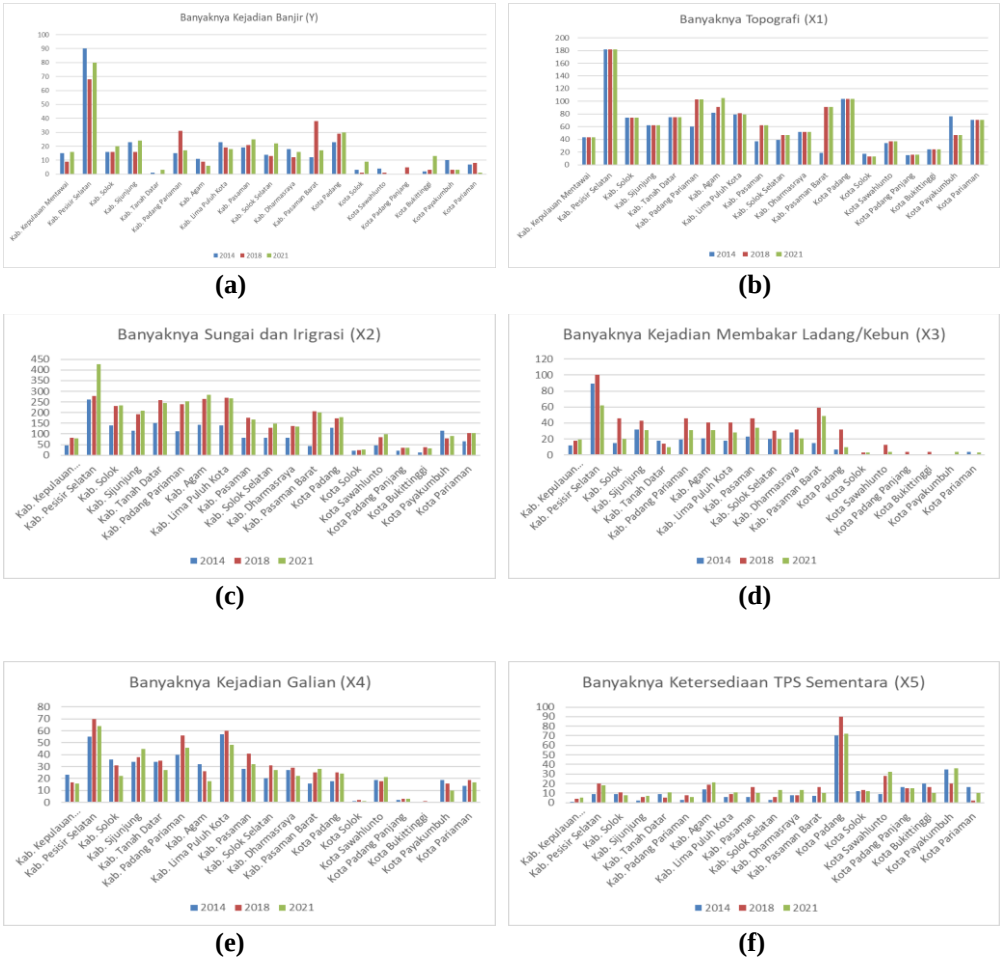
In this research, analysis was carried out using R and STATA software. The steps for carrying out the analysis in this research are as follows.

1. Describe the general picture of the number of flood events in West Sumatra.
2. Carry out a multicollinearity test
3. Carry out modelling using the Poisson regression model with the Generalized Estimating Equation (GEE) method with the following steps:
   a. Carry out significance tests of model parameters
   b. Calculate the Quasi Likelihood under Independence Model Criterion (QIC) value for each model based on the Working Correlation Structure (WCS).
   c. Estimating the parameters of the Poisson regression model based on the best WCS
4. Because there is overdispersion, proceed with modeling using a negative binomial regression model using the Generalized Estimating Equation (GEE) method with the following steps.
   a. Carry out significance test of model parameters
   b. Calculate the Quasi Likelihood under Independence Model Criterion (QIC) value for each model based on the Working Correlation Structure (WCS).
   c. Estimating the parameters of the Poisson regression model based on the best WCS
5. Determine the best model based on the lowest Quasi Likelihood under Independence Model Criterion (QIC) value of the two regression models used, which are Poisson regression and negative binomial regression

3. RESULTS AND DISCUSSION

3.1 Descriptive Statistics

A general description of each variable based on factors causing flooding in West Sumatra Province in 2014, 2018 and 2021 is presented in **Figure 1** as follows:



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Figure 1. Diagram of Each Variable, (a) Number of Flood Events, (b) Amount of Topography, (c) Number of Rivers and Irrigation, (d) Number of Instances of Burning Fields/Crops, (e) Number of Excavation Events, (f) Number of Temporary Dumping Sites Available

In summary, presented in **Table 3** as follows:

Table 3. Descriptive Statistics for Each Variable

Variable	Max	District/Citty	Min	District/City	Average
y	90	South Coast District	0	Tanah Datar District	16.2807
x_1	182	South Coast District	13	Solok City	65.03509
x_2	428	South Coast District	14	Bukittinggi City	140.7544
x_3	100	South Coast District	0	Bukittinggi City	22.35088
x_4	70	South Coast District	0	Bukittinggi City	26.12281
x_5	90	Padang City	1	Mentawai Islands	15.5614

3.2 Multicollinearity Test

The multicollinearity test is a test used to show the existence of a correlation or strong relationship between two or more independent variables in a multiple linear regression [14]. A good regression model should not have high correlation between independent variables. The multicollinearity test was carried out by looking at the Variance Inflation Factor (VIF). The multicollinearity test hypothesizes are as follows:

H_0: There is no multicollinearity
H_1: There is multicollinearity

Table 4. Multicollinearity test result

Variable	VIF Score
x_1	5.43
x_2	4.05
x_3	3.17
x_4	3.72
x_5	1.49

Based on **Table 4** It can be decided that it failed to reject H_0 because the VIF value is below 10, so it can be concluded that there is no multicollinearity in the data.

3.4 Poisson Regression Modeling with Generalized Estimating Equation (GEE)

Based on the multicollinearity test, it shows that there is no multicollinearity in each variable so it is feasible to carry out Poisson regression modeling with Generalized Estimating Equation (GEE). In determining the optimal model of Poisson regression, it can be determined by the Quasilikelihood Index Criterion (QIC) value **[15]**. The model that has the lowest QIC value for Working Correlation Structure (WCS) is considered the best model. The results of QIC for each WCS with Poisson regression are presented in **Table 5.**

Table 5. QIC results of each WCS with Poisson Regression

Working Correlation Structure (WCS)	QIC value
Exchangeable	287.627
Independent	263.365
Unstructured	275.710

Based on **Table 5**, it shows that the GEE model with the "Independent" type of WCS gives the lowest QIC value of 263.365, despite the differences being relatively small between all the WCS formed. This shows that the most optimal GEE model is the GEE model with "Independent" type WCS. Next, significant parameter testing was carried out simultaneously. Simultaneous parameter testing was carried out using a $D(\widehat{\boldsymbol{\beta}})$ value of 263.02 with degrees of freedom of 57. With hypothesis as follows:

H_0 : All parameters do not have a significant effect in the model or $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_1: There is at least one parameter that has a significant influence in the model or at least one $\beta_j \neq 0$, for $j = 1, 2, 3, 4, 5$

The decision to reject H_0 is if $D(\widehat{\boldsymbol{\beta}}) > \chi^2_{(0,05;57)}$ with a significance level of 5% so that the value of $\chi^2_{(0,05;57)} = 216{,}44$ is obtained. Therefore, a decision can be obtained to reject H_0 because the value of $D(\widehat{\boldsymbol{\beta}}) > \chi^2_{(0,05;57)}$.. Thus, it can be concluded that there is at least one parameter that has a significant influence in the model. The results of parameter estimation in this study using Poisson regression based on "Independent" WCS are presented in **Table 6.**

Table 6. GEE Parameter Estimation Results with Poisson Regression Based on Independent Type WCS

Parameter	Estimation	$\boldsymbol{p-value}$
$\boldsymbol{\beta_0}$	1.517472	0.008
$\boldsymbol{\beta_1}$	0.0048048	0.929
$\boldsymbol{\beta_2}$	-0.0000571	0.000
$\boldsymbol{\beta_3}$	0.0129333	0.000
$\boldsymbol{\beta_4}$	0.0126577	0.003
$\boldsymbol{\beta_5}$	0.0064154	0.000
QIC = 263.365	**Deviance = 263.02**	**Degree of Freedom = 57**

Based on **Table 6**, it is shown that the results of the estimated Poisson regression parameters show that the variables x_1, x_3, x_4, x_5 with a significance level of 0.05 have an influence on the occurrence of floods in each district/city in West Sumatra. β_0 or intercept also has a significant influence on the occurrence of floods in each district/city in West Sumatra, meaning that if all dependent variables are taken into account or not, there will still be a positive tendency for floods to occur for each additional year. It can be concluded that there is a tendency of increase in the frequency of floods in each district/city in West Sumatra every year. Therefore, the model obtained based on these results is as follows:

$$(\hat{\mu}) = \exp\left(1.517472 - 0.0000571x_2 + 0.0129333x_3 + 0.0126577x_4 + 0.0064154x_5\right)$$

3.5 Overdispersion

In the Poisson regression test, the QIC value based on the "Independent" WCS was 263.365, meaning despite being the least compared to the other types, the QIC value was still very large. Apart from that, a deviance value of 263.02 was obtained with degrees of freedom of 57 and a dispersion value of 4.6144. Aside from being obtained from the software results, the dispersion value can also be calculated using the deviance value divided by the degrees of freedom. Because the dispersion value is greater than 1, it can be concluded that there is overdispersion, so that overall the Poisson model that has been obtained cannot be used because the equidispersion assumption is not met and causes the model results to be less accurate. One strategy to overcome overdispersion in Poisson regression is to replace the Poisson distribution with another distribution that is more flexible [16]. In this case, the alternative distribution used is the negative binomial distribution. The negative binomial regression approach was chosen because the Poisson distribution can be considered as a special case of the negative binomial distribution with parameter $\alpha = 0$ [17].

3.6 Negative Binomial Regression Modelling with Generalized Estimating Equation (GEE)

Negative binomial regression was used to overcome overdispersion in Poissson regression [18]. Next step is to determine the best model from the types of Working Correlation Structure (WCS). In determining the best model for negative binomial regression, it can be seen from the Quasilikelihood Index Criterion (QIC) value. The model that has the lowest QIC value for WCS is considered the best model. The results of QIC for each WCS with Poisson regression are presented in Table 7.

Table 7. QIC Results for each WCS using Negative Binomial Regression

WCS	QIC Value
Exchangeable	41.345
Independent	39.513
Unstructured	40.524

Based on Table 7, it is shown that the GEE model with the "Independent" WCS gives the minimum QIC value of 39.513, despite the differences being relatively small between all the WCS formed. This shows that the best GEE model is the GEE model with "Independent" type WCS. Next, significant parameter testing was carried out simultaneously. Simultaneous parameter testing was carried out using a $D(\hat{\boldsymbol{\beta}})$ value of 34.61 and degrees of freedom of 57. The hypothesizes are as follows:

H_0 : All parameters do not have a significant effect in the model or $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_1: There is at least one parameter that has a significant influence in the model or at least one $\beta_j \neq 0$, for $j = 1, 2, 3, 4, 5$

The decision to reject H_0 is if $D(\hat{\boldsymbol{\beta}}) > \chi^2_{(0,05;57)}$ with a significance level of 5% so that the value of $\chi^2_{(0,05;57)} = 20.22$ is obtained. Therefore, a decision can be obtained to reject H_0 because the value $D(\hat{\boldsymbol{\beta}}) > \chi^2_{(0,05;57)}$. Thus, it can be concluded that there is at least one parameter that has a significant influence on the model. The results of parameter estimation in this study using negative binomial regression based on "Independent" WCS are presented in Table 8.

Table 8. GEE Parameter Estimation Results with Negative Binomial Regression Based on Independent Type WCS

Parameter	Estimation	$\boldsymbol{p-value}$
$\boldsymbol{\beta_0}$	1.383329	0.000
$\boldsymbol{\beta_1}$	0.0042941	0.608
$\boldsymbol{\beta_2}$	-0.0025449	0.412
$\boldsymbol{\beta_3}$	0.0234557	0.041
$\boldsymbol{\beta_4}$	0.203363	0.195
$\boldsymbol{\beta_5}$	0.0085826	0.397
QIC = 39.513	***Deviance* = 34.61**	**Degree of Freedom = 57**

Based on Table 8, it is shown that the results of the negative binomial regression parameter estimation showed that only one variable with a significance level of 0.05 had an influence on the occurrence of floods in each district/city in West Sumatra, which is the variable x_3, fields cleared through burning.

3.7 Best Correlation Selection

The selection of the best model is done by looking at the lowest QIC value. A summary of the two models used, namely Poisson regression, and negative binomial regression with significant variables is presented in **Table 9**.

Table 9. Best Model Selection

Regression Model	QIC Value
Poisson	263.365
Negative Binomial	39.513

Based on **Table 9**, it shows that the negative binomial regression model using the GEE method gives the lowest QIC value, so this model is the best model.

4. CONCLUSIONS

The data used in the research meets the multicollinearity assumption, but because there is more than one overdispersion, the distribution function family that can be used to overcome this is negative binomial. The results of parameter estimation using negative binomial regression showed that only one variable was significant, which is the number of incidents of field burning (x_3). The regression equation of the negative binomial regression model is as follows:

$$y = \exp(1.383329 + 0.0234557x_3)$$

From the two regression models used, which are Poisson and negative binomial, it was found that the negative binomial regression model was the best model because it had the lowest QIC value of 39.51.

This study has several limitations such as inconsistent flood data quality, missing data, and changes in measurement methods, inadequate testing of the correlation structure, non-optimal survey variables, and limited generalizability of the model. Future research should improve data quality, explore flexible correlation structures, use alternative models like mixed-effects models, and include variables such as land use changes and human activities. Comparative studies and interdisciplinary approaches can also provide more comprehensive insights.

REFERENCES

[1] S. Hardiyanto and D. Pulungan, "Komunikasi Efektif Sebagai Upaya Penanggulangan Bencana Alam di Kota Padangsidimpuan.," *Jurnal Interaksi: Jurnal Ilmu Komunikasi,* pp. 3(1), 30-39, 2019.

[2] A. Rosyidie, "Banjir: fakta dan dampaknya, serta pengaruh dari perubahan guna lahan.," *Jurnal perencanaan wilayah dan kota,* pp. 24(3), 241-249, 2013.

[3] D. Ramadhan, "Antara News," BNPB: Tiga wilayah di Sumatera Barat dilanda banjir dan longsor, 14 Juli 2023. [Online]. Available: https://www.antaranews.com/berita/3635613/bnpb-tiga-wilayah-di-sumatera-barat-dilanda-banjir-dan-longsor#mobile-src.

[4] Antara, "tempo.co," Ribuan Warga di Wilayah Ini Terdampak Banjir hingga Mengungsi, 20 November 2023. [Online]. Available: https://tekno.tempo.co/read/1798883/ribuan-warga-di-wilayah-ini-terdampak-banjir-hingga-mengungsi.

[5] panda, Menghadapi Bencana Banjir: Upaya Pengelolaan dan Mitigasi Risiko untuk Masyarakat Desa, 01 06 2024. [Online]. Available: https://www.panda.id/menghadapi-bencana-banjir-upaya-pengelolaan-dan-mitigasi-risiko-untuk-masyarakat-desa/. [Accessed 19 08 2024].

[6] A. Setiawan, "Regresi Poisson Menggunakan Generalized Estimating Equation (Studi Kasus: Data Longitudinal Frekuensi Terjadinya Banjir Di Jawa Timur Tahun 2011- 2013).," in *Master's Thesis, Institut Teknologi Sepuluh Nopember*, Surabaya, 2017.

[7] L. Wu, Mixed Effect Models for Complex data, New York: CRC Press, Taylor and Francis Group, 2010.

[8] R. Naelu, "Karakteristik Penduga Parameter Generalized Estimating Equation (GEE) pada Data Longitudinal," in *Skripsi, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung*, Lampung, 2016.

[9] X. Liu, Methods and applications of longitudinal data analysis, London: Elsevier , 2015.

[10] M. Wang, L. Kong, Z. Li and L. Zhang, "Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples," *Statistics in medicine,* vol. 35, no. 10, pp. 1706-1721, 2016.

[11] Danardono, Analisis Data Longitudinal, Yogyakarta: Gadjah Mda University Press., 2015.

[12] T. W. Utami, "Analisis regresi binomial negatif untuk mengatasi overdispersion regresi poisson pada kasus demam berdarah dengue," *Jurnal Statistika Universitas Muhammadiyah Semarang*, vol. 1, no. 2, 2013.

[13] D. E. Ashari, "Faktor-Faktor yang Mempengaruhi Banyaknya Pneumonia Balita Di Jawa Timur Menggunakan Generalized Poisson Regression (GPR) dan Negative Binomial Regressiion (NBR)," *Institut Teknologi Surabaya*, 2014.

[14] N. Shrestha, "Detecting multicollinearity in regression analysis," *American Journal of Applied Mathematics and Statistics*, vol. 8, no. 2, pp. 39-42, 2020.

[15] A. Crisci, L. D'Ambra and V. Esposito, "A generalized estimating equation in longitudinal data to determine an efficiency indicator for football teams," *Social Indicators Research*, vol. 146, no. 1-2, pp. 249-261, 2019.

[16] L. H. Vanegas and L. M. Rondon, "A data transformation to deal with constant under/over-dispersion in Poisson and binomial regression models," *Journal of Statistical Computation and Simulation*, vol. 90, no. 10, pp. 1811-1833, 2020.

[17] P. A. Omar and T. M. Hussian, "Fitting of Generalized Poisson Regression and Negative Binomial Regression models for Analyzing of Count Time Series Event," *Polytechnic Journal of Humanities and Social Sciences*, vol. 4, no. 2, pp. 116-126, 2023.

[18] Y. Setyawan, K. Suryowati and D. Octaviana, "Application of Negative Binomial Regression Analysis to Overcome the Overdispersion of Poisson Regression Model for Malnutrition Cases in Indonesia," *Parameter: Journal of Statistics*, vol. 2, no. 2, pp. 1-9, 2022.