JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 10, No. 1, January 2026, pp. 71-83

The Probability Model of Earthquake Frequency in the Enggano Segment Using Poisson Mixture Models

Siska Yosmar[1*], Ramya Rachmawati[1], Septri Damayanti[1], Jose Rizal[1]
[1]Department of Mathematics, University of Bengkulu, Indonesia
siskayosmar@unib.ac.id

Article History:
Received : 25-07-2025
Revised : 04-10-2025
Accepted : 08-11-2025
Online : 01-01-2026

Keywords:
AIC;
BIC;
Earthquake;
Poisson Mixture Models;
Poisson Hidden Markov Models.

ABSTRACT

An earthquake is a natural disaster that occurs suddenly resulting in numerous casualties, such as loss of life and property. Bengkulu Province is among the provinces affected by severe earthquakes. Studies on probability models for the frequency of earthquake events in Bengkulu Province are still scarce, as outlined in the 2017 book "Map of Sources and Hazards of Indonesian Earthquakes." This research uses Poisson mixture models to build a probability model for the frequency of earthquake events in the Enggano segment, located in the coastal area of Bengkulu Province. ..The phases of model building are the model diagnosis phase, testing the dispersion state relative to the Poisson distribution, testing the dependence of research data on time variables using the Ljung-Box test, and testing the criteria for selecting the best model using the Bayesian Tests Measures of Information Criterion (BIC) and Akaike Information Criterion (AIC). Annual earthquake frequency data from January 1, 1971, to December 31, 2022, were retrieved from the USGS catalog of data on the frequency of major earthquakes with a magnitude of Mw ≥ 4.40, which occurred a total of 633 times. After completing the model building phase, the AIC and BIC values for each model were determined by determining the number of unobserved groups. Both Poisson mixture models and Poisson hidden Markov models produced the same number of unobserved groups of 3 groups with AIC=302.91 and BIC=324.38.

https://doi.org/10.31764/jtam.v10i1.33446

This is an open access article under the CC–BY-SA license

A. INTRODUCTION

Earthquakes are a form of natural disaster that can cause significant casualties and economic losses (Reid, 2015). The occurrence of earthquakes is caused by the movement of tectonic plates (an unobservable process) that collide due to the mechanism of convection currents from the Earth's core (Boden, 2016). Based on the mechanism, there are three types of movement from one tectonic plate to another convergent, divergent, and transformative. Indonesia is an earthquake-prone area. This movement is because, geologically speaking, Indonesia's territory lies at the meeting of three main tectonic plates in the Pacific Ring of Fire, namely the Indo-Australian Plate, the Eurasian Plate, and the Pacific Plate (Irsyam et al., 2020). Due to the strong suspicion that there is a mutual connection between motion mechanisms and plate tectonics, this topic is interesting from various perspectives, such as spatial analysis, seismic analysis, and mathematical modelling.

Earthquake engineering researchers who study the relationship between earthquake events and plate tectonic dynamic activity from a mathematical modeling perspective include (Orfanogiannaki & Papadopoulos, 2014; Yip et al., 2017; Rizal et al., 2018; Rizal et al., 2023). In

summary, several comments are based on the results of a review of the four articles. The first clue refers to the seismic data used, where the process of earthquake events (observed process) is represented by seismic data in the form of earthquake events with a magnitude of $Mw \geq Mc$ (Magnitude of Completeness). In contrast the dynamic process of tectonic plates (unobserved process) is presented in the form of categories or levels, e.g., low, medium, high dynamics or other categories that describe the behavior of the dynamics of tectonic plates. The second note refers to the probability model applied, namely a mixture model of the Poisson distribution with a hidden Markov model (MMT). The MMT model is applied to process parameters from unobserved state spaces that are assumed to satisfy Markov properties, while the Poisson distribution is applied to state-dependent processes in unobserved states whose state space can be observed (Zucchini et al., 2017). The third reference to the algorithm for estimating the parameters of the Poisson MMT model is the EM (Expectation Maximization) algorithm (Dempster et al., 1977). The fourth note relates to the fundamental difference between the five researchers, which lies in the magnitude limit used in calculating earthquake events. This magnitude limit is determined using the Mc value from the selected catalog data and the area used as a research object.

In this research, the Enggano segment, which lies in the Sumatra subduction zone, was selected as the research area. Our motivation to choose this area was based on the results of studies by Sieh et al. (2007) and McCloskey et al. (2008), where after the major earthquake in Aceh-Andaman (December 26, 2004, $M_w = 9.2$) and in Nias-Simeuleu (March 25, 2005, $M_w = 8.7$), the next major earthquake will be around the Mentawai and Enggano Islands with magnitude $M_w \geq 8$ predicted. The results of this research are confirmed by the geodetic and paleoseismic calculations carried out by Sieh et al. (2008); Chlieh et al. (2008). The results of this research show a large earthquake risk with the formation of tsunami waves around the Mentawai and Enggano island segments. However, studies on earthquake probability models in these two segments are still scarce, as explicitly stated in the articles (Irsyam et al., 2017; Rizal et al. 2022).

Zucchini et al. (2017) explain that the probability model for the number of occurrences of an event (e.g., the frequency of earthquakes) can be viewed as a Poisson process that follows the Poisson distribution, with the characteristics of the sample mean and variance is equal. However, during implementation, overdispersion conditions regarding the Poisson distribution often arise, namely that the sample mean is greater than the variance. One of the factors causing overdispersion is the presence of data groups that are not observed in the modeled population data. One method to overcome the overdispersion problem is to use a mixture model. Furthermore, based on the nature of the dependence of these unobserved data sets, mixture models can be classified into two types, namely independent Poisson mixture models (PMMs) and dependent Poisson mixture models (PHMMs). Explanations of these two models can be found in the next two sections.

The paper is organized as follows: In Section methods, we explain the data and the models used in our work. A general description of PMMs and PHMMs, including the EM algorithm for estimating the model's parameters, is presented in two subsections. In the section Results and Discussion, we report the results of the data analysis and some discussions concerning the

relevancy of the present results with other studies. In the last section, the conclusion and suggestions are written.

B. METHODS

Before explaining the models used in this research, namely PMMs and PHMMs, we first explain the probability function model of Poisson distribution. The theoretical probability function of the Poisson distribution can be described as follows: Assume X_t is a random variable following the Poisson distribution with the parameter $\lambda_t > 0$, written as $X_t \sim Poi(\lambda_t)$, then the probability function of X_t is written as follows:

$$p(x_t; \lambda_t) = \frac{e^{(-\lambda_t)}(\lambda_t)^{x_t}}{x_t!}, x_t = 0,1,2,.. \quad (1)$$

The Poisson distribution is characterized by having the same sample mean and variance. However, when using them, the properties of the Poisson distribution are often violated. If the variance value is greater than the average, this condition is called the condition of relative overdispersion in the Poisson distribution. One approach to overcome this problem of relative overdispersion is to use a mixture model. Mixture models are designed to account for the nature of data heterogeneity due to the existence of groups of unobserved data that are part of the main data. In addition, if the data groups are identified as independent, which shows from the test results that the main data does not depend on time, then the mixture model of these data groups is called a Poisson distribution independent mixture model or written as an independent mixture (Zucchini., et al. (2017).

Independent Mixture Models generally consist of a finite number of groups (m components) and a Poisson distribution. Let $\delta_1, \delta_2, \delta_3, \dots, \delta_m$ be the probabilities in each m-group and let $p(1), p(2), p(3), \dots, p(m)$ be the probabilities of the density function for each group. For a discrete random variable X_t indicating that the random variable has a mixed Poisson distribution, it is formulated as follows:

$$p(x_t; \lambda_t) = \sum_{i=1}^{m} \delta_i p_i(x_t) \quad (2)$$

The groups described in equation (2) have the property of being independent of each other. But in reality, these groups may be interdependent. In this case, Poisson Hidden Markov Models (PHMMs) can be used as an alternative model. Hidden Markov Models (HMMs) are stochastic processes that consist of two parts. The first part is the unobserved part of the process $\{X_t, t \in \mathbb{N}\}$, which is assumed to satisfy Markov properties. The second part is the observation process $\{C_t, t \in \mathbb{N}\}$. This process is based on hidden states. The state space X_t depends only on the current state C_t and not on previously observed states X_{t-1}. The hidden Markov model $\{X_t, t, \mathbb{N}\}$ is a mixed distribution that depends on $X^{(t)}$and $C^{(t)}$and represents past events from time 1 to time t, which can be concluded from this simple model with the following equation:

$$\Pr\left(\left(C_t|C^{(t-1)}\right) = \Pr\left(C_t|C_{(t-1)}\right)\right), t = 2,3,4, \ldots \quad (3)$$

and

$$\Pr\left(X_t|X^{(t-1)}, C^t\right) = \Pr(X_t|C_t), t \in \mathbb{N} \quad (4)$$

If the Markov chain $\{C_t\}$ has m hidden states, it can be concluded that $\{X_t\}$ are PHMMs with m states. A commonly used method for estimating parameters in HMMs is the Estimation Maximization Algorithm (EM algorithm) method (Dempster et al., 1977). In the context of HMMs, the EM algorithm is often referred to as the Baum-Welch algorithm, where the Markov chain in HMMs is homogeneous and does not have to be stationary. The parameters of the HMMs estimated by the EM algorithm are the dependent state distribution pi, the transition probability matrix Γ and the initial distribution δ. In its application, the EM algorithm requires tools, namely forward chances and backward chances, both of which can be used to predict states (Zucchini et al. 2017). The forward probability α_t for $t = 1,2, \ldots, T$ is defined as a row vector:

$$\alpha_t = \delta P(x_i)\Gamma\mathbf{P}(x_2) \ldots \Gamma\mathbf{P}(x_t) = \boldsymbol{\delta}\mathbf{P}(x_1)\prod_{s=2}^{t}\Gamma\mathbf{P}(x_s) \quad (5)$$

The value δ is the initial distribution of the Markov chain. Based on the definition of forward odds in equation (5), for $t = 1, 2, \ldots, T - 1$. Next is the backward probability β_t for $t = 1, 2, \ldots, T$ which is defined as a row vector:

$$\beta_t = \Gamma\mathbf{P}(x_{t+1})\Gamma\mathbf{P}(x_{t+2}) \ldots \Gamma\mathbf{P}(x_T)\mathbf{1}' = \left(\prod_{s=t+1}^{t}\Gamma\mathbf{P}(x_s)\right)\mathbf{1}' \quad (6)$$

where for the value $t = T, B_T = 1$. From equations (5) and (6), we get :

$$\Pr\left(C_t = j|X^{(T)}\right) = x^{(T)} = \frac{\alpha_t(j)\beta_t(j)}{L_T} \quad (7)$$

and

$$\Pr(C_{t-1} = j, C_t) = k|(X^{(T)} = x^T) = \frac{\alpha_{t-1}(j)\gamma_{jk}p_k(x_t)\beta_t(k)}{L_T} \quad (8)$$

In HMMs, especially where the sequence of Markov chain states is not observed, there may be missing data (missing values) in the sequence, resulting in incomplete data. The EM algorithm is an iterative method used to calculate the maximum likelihood estimate for incomplete data, thus obtaining complete log-likelihood data. In each iteration of the EM algorithm, there are two stages, namely the expectation stage or E stage (E-step) and the maximization stage or M stage (M-step). After estimating the parameters of several possible models to be applied, the model that best fits the modeled earthquake data is determined. In this study, we used two criteria in selecting the best model, namely the Akaike Information Criterion (AIC) and the Bayesian Information Criteria (BIC) (Kuha 2004). The formula for these

two criteria is explained in the next paragraph. The formula for AIC can be calculated using the equation:

$$\text{AIC} = 2\text{K} - 2\log(\text{Maximum Likelihood}) \tag{9}$$

The BIC value is now formulated as follows:

$$\text{BIC} = -2\log(\text{Maximum Likelihood}) + \text{k}\log(\text{n}) \tag{10}$$

where n is the amount of data and k is the number of parameters to be estimated. As a reminder, the best model is determined using the smallest AIC and BIC values.

C. RESULT AND DISCUSSION

The seismic data used in this study are annual earthquake frequency data that occurred in the Enggano segment with observation periods from January 1, 1971, to December 31, 2022. These data were obtained from the United States Geological Survey (USGS) Earthquake Data Catalog. It contains three types of earthquake events, namely pioneer, main, and aftershock events, a total of 11,613 earthquake data, of which 633 are main earthquake events. Using this data, a process was then carried out to tabulate the frequency of earthquake events with a magnitude $M_w \geq 4.40$. Descriptive statistics from research data are briefly presented below:

Table 1. Descriptive statistics on the number of earthquake events

Mean	Median	Variance	Maks	Min	Skewness	Kurtosis
12,17	12,00	38,18	26	2	0,15	2,21

Based on Table 1, two aspects can be explained as follows. The first aspect refers to the comparison between the variance value and the mean, where the variance value is 38,18 and the mean is 12,17, so it can be concluded that the variance value is larger than the mean. In other words, there is an overdispersion of the Poisson distribution in the data on the frequency of earthquakes in the Enggano segment. In addition, the second aspect is related to the skewness value and the kurtosis value, where the skewness value is greater than 0 and the kurtosis value is less than 3. From these two values, it follows that the data on the frequency of earthquake events in the Enggano segment is not distributed symmetrically or more precisely as a function of odds and tends to dominate on the right side. Based on the results of descriptive statistical analysis of the data, it can be concluded that the data on earthquake events in the Enggano segment are overdispersed.

One way to overcome the condition of overdispersion of data is to apply a mixed model (independent and dependent) of the Poisson distribution. The selection of the two models is adapted to the properties of the data. If the research data depends on time, the chosen model is a dependent mixed model, using Poisson Hidden Markov Models (PHMMs). While the data is not dependent on time, the Poisson Mixture Models (PMMs) is preferred. In this research, the research method for testing the data dependence on time uses the Ljung-Box test (Ljung and Box 1978).

The Ljung-Box test is performed to check whether the modeled data is time-dependent. For time-dependent data, the $p - value < \alpha$; otherwise the data is not time-dependent if the $p -$

$value > \alpha$. In this study, the α used was 0,05. The Ljung-Box test hypothesis is as follows: $H_0: \rho^2(h) = 0$, which means that the data does not depend on time, while $H_1: \rho^2(h) \neq 0$, which means that the Data depends on time. The statistics suggested by Ljung and Box are:

$$Q = n(n+2)\sum_{k=1}^{K}\left(\frac{r_k^2}{n-k}\right) \tag{11}$$

The description of Equation (11) is as follows: n is the number of observation data, K is the number of selected delays, and r_k^2 is the sample correlation. The test criteria for the Ljung-Box test are: If the hypothesis is tested: $p - value < \alpha$, then H_0 is rejected. The *p-value* obtained from the Ljung-Box test for the modeled data is $2{,}02 \times 10^{-5}$. This value is smaller than the selected α value, so it can be concluded that the data on the frequency of earthquake events in the Enggano segment is time-dependent. Based on these results, the Poisson distribution mixture model is the more suitable distribution model.

In the implementation phase of the Poisson distribution mixture model, a trial-and-error approach (repetitive method) is generally used from different possible states. In this study, the best model was sought from three models. The three models are hidden/unobserved state models (m), namely (2, 3, 4). The criteria for selecting a multistate model are based on the smallest BIC and AIC values. The next step to determine the input parameters is to take the value $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)$, namely the parameter value for the average number of earthquake events per year, and the initial probability of occurrence $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)$. The first step in finding the parameter values is to create a frequency distribution table from data on the number of earthquakes, where the number of classes in the table is determined by the number of hidden conditions.

In this study, values ranging from 2 to 26 for the number of earthquakes were used to divide the intervals for each class using Rstudio. In this article, we will only explain the case m=3. Given three hidden states with an average number of earthquakes $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$, then there are 3 classes with interval lengths for each class. Calculate the value $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ by inputting earthquake frequency data into each hidden state group based on the sample space. Parameter $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3)$ is the initial probability from hidden state 1 to hidden state 3 obtained by calculating the number of frequencies in each group divided by the total frequency of the hidden state. The estimated parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\delta}$ for the case of 3 hidden states can be seen in the following Table 2.

Table 2. Results of PMM parameter calculation for hidden conditions in case 3

Amount	Frequency	Hidden Events	$\boldsymbol{m = 1}$	$\boldsymbol{m = 2}$	$\boldsymbol{m = 3}$
1	6	1	6	-	-
2	7	1	7	-	-
3	7	1	7	-	-
4	7	1	7	-	-
5	8	1	8	-	-
6	8	1	8	-	-
7	8	1	8	-	-
8	9	1	9	-	-

Amount	**Frequency**	**Hidden Events**	$m = 1$	$m = 2$	$m = 3$
9	9	1	9	-	-
10	9	1	9	-	-
⋮	⋮	⋮	⋮	⋮	⋮
43	18	3	-	-	18
44	18	3	-	-	18
45	19	3	-	-	19
46	19	3	-	-	19
47	21	3	-	-	21
48	22	3	-	-	22
49	22	3	-	-	22
50	22	3	-	-	22
51	24	3	-	-	24
52	26	3	-	-	26
	f	52	18	9	25
	λ		9,11	3,22	17,60
	δ		0,35	0,17	0,48

After obtaining the estimated values of $\boldsymbol{\lambda}$ and $\boldsymbol{\delta}$ over a period per year, it can be declared that the first hidden state has an average number of earthquakes of 9,11 events with an initial occurrence probability of 0,35 for the average occurrence of the second hidden state of 3,22 with an initial probability of 0,17 and for the average occurrence in the third hidden state of 17,60 with an initial probability of 0,48. Next, an iteration process is performed to obtain convergent model parameter values. The results of Poisson Mixture Models modeling for multiple values of m are shown in Table 3.

Table 3. Poisson Mixture Model Results on Earthquake Data

Model	$\boldsymbol{i}$	$\boldsymbol{\lambda}$	$\boldsymbol{\delta}$	***-llk***	**Iteration**	**AIC**	**BIC**
$m = 1$	1	12.17	1	197.62	2	397.24	399.19
$m = 2$	1	15.42	0.63	167.63	22	341.26	347.11
	2	6.53	0.37				
$m = 3$	1	9.11	0.35	165.33	55	**339.66**	**340.41**
	2	3.22	0.17				
	3	17.60	0.48				
$m = 4$	1	5.53	0.33	165.32	89	344.66	358.32
	2	11.38	0.31				
	3	16.92	0.25				
	4	22.83	0.12				

Based on a comparison of the AIC and BIC values for the m values that were tested, the smallest value was in the model for m=3 with an AIC value = 339.66 and a BIC value = 340.41. Furthermore, research data modeling using PHMMs is also studied, where the parameters are estimated using the EM algorithm. The following are the steps taken along with the processed results to obtain the three models for estimating the number of earthquakes and determining the best model.

At this stage of determining the input parameters, we calculate the value $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)$, namely the parameter value of the average number of earthquake events per year, with the initial probability of the event $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)$ and the hidden state transition probability matrix $\boldsymbol{\Gamma}$ size $(n\ x\ n)$. The initial step finding the parameter values is to create a frequency distribution table from data on the number of earthquakes with the number of classes in the table determined by the number of hidden conditions that will be provided. In this study, values in the range 2 to 26 for the number of earthquakes were taken to divide the intervals into each class uniformly. For example, in a model with m=3, if given three hidden states with an average number of earthquakes $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$, then there are three classes with interval lengths for each class.

$$c = \text{minimum value} + \frac{\text{range}}{\text{many classes}} = 2 + \frac{24}{3} = 10$$

Based on the c value above, for the sample space the number of earthquakes in hidden state 1 is {2, 3, 4, …, 10} out of {11, 12, 13, …, 18} in hidden state 2 and {19, 20, 21, …, 26} enter the hidden state 3. To calculate the value $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ is to enter earthquake frequency data into each group of hidden states based on the sample space. Parameter $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3)$ is the initial probability of hidden state 1 to hidden state 3, obtained by calculating the number of frequencies in each group divided by the overall frequency of hidden states. The results of calculating the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\delta}$ for the case of 3 hidden states can be seen in the table below:

Table 4. Results of PHMMs parameter calculation for hidden conditions in case 3

Amount	**Frequensi**	**Hidden events**	$\boldsymbol{m = 1}$	$\boldsymbol{m = 2}$	$\boldsymbol{m = 3}$
1	2	1	2	-	-
2	2	1	2	-	-
3	2	1	2	-	-
4	3	1	3	-	-
5	4	1	4	-	-
6	3	1	3	-	-
7	3	1	3	-	-
8	7	1	7	-	-
9	8	1	8	-	-
10	8	1	8	-	-
⋮	⋮	⋮	⋮	⋮	⋮
42	14	2	-	14	-
43	14	2	-	14	-
44	24	3	-	-	3
45	14	2	-	2	-
46	10	1	1	-	-
47	19	3	-	-	19
48	22	3	-	-	22
49	21	3	-	-	21
50	17	2	-	17	-
51	15	2	-	15	-
52	22	3	-	-	22
	f	62	23	21	8

Amount	**Frequensi**	**Hidden events**	$\boldsymbol{m=1}$	$\boldsymbol{m=2}$	$\boldsymbol{m=3}$
	λ		6,00	14,24	19,25
	δ		0,44	0,40	0,15

After obtaining the values of $\boldsymbol{\lambda}$ and $\boldsymbol{\delta}$, t was found that the first hidden state had an average number of earthquakes of 6,00 events with an initial probability of occurrence of 0,44, for the average occurrence in the second hidden state of 14,24 with a The initial probability is 0,40 and for the average event in the third condition is 19,25, where the initial probability of the event is 0,15. Parameter value $\boldsymbol{\Gamma}$, namely the probability matrix for the hidden state transition, where the elements in the matrix are obtained by calculating the frequency of each possible hidden state transition, which is then divided by the total number of each row in the hidden state. There are three possible hidden state transitions, shown in the following Table 5:

Table 5. Probabilities in 3 hidden states

Unobservable state	**1**	**2**	**3**
1	17/23	4/23	2/23
2	5/21	12/21	4/21
3	0/7	5/7	2/7

so we get the transition probability matrix for the three hidden states as follows:

$$\boldsymbol{\Gamma} = \begin{bmatrix} 0.74 & 0.17 & 0.09 \\ 0.24 & 0.57 & 0.19 \\ 0 & 0.71 & 0.29 \end{bmatrix}$$

Based on the above transition probability matrix, it can be interpreted that if this period is in the first hidden state, the probability that the frequency of future earthquakes will be in the first hidden state is 0.74. When this period is in hidden state 1, the probability that the frequency of earthquakes in the future will be in hidden state 2 is 0.17, and when it is in hidden state 1, the probability that the frequency of earthquakes in the future will be like this in the hidden state 3 is 0.09 and so on. The results of the PHMM modelling can also be seen in Table 6.

Table 6. Input parameters λ, δ and Γ for each PHMMs

$\boldsymbol{m}$	$\boldsymbol{i}$	$\boldsymbol{\lambda}$	$\boldsymbol{\delta}$	$\boldsymbol{\Gamma}$			
				1	**2**	**3**	**4**
$m=2$	1	8.219	0.615	0.750	0.250	-	-
	2	18.500	0.385	0.368	0.632	-	-
$m=3$	1	6.000	0.442	0.739	0.174	0.087	-
	2	14.238	0.404	0.238	0.571	0.190	-
	3	19.250	0.154	0.000	0.714	0.286	-
$m=4$	1	5.529	0.327	0.588	0.235	0.118	0.059
	2	11.375	0.308	0.313	0.438	0.125	0.125
	3	16.923	0.205	0.077	0.308	0.462	0.154
	4	22.833	0.115	0.000	0.200	0.600	0.200

Next, perform an iterative process on the initial values of the parameters $\hat{\boldsymbol{\lambda}}$, $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\Gamma}}$ for each model using the EM (Expectation Maximization Algorithm) algorithm. After the parameter estimation results are obtained, the next step is to compare the AIC and BIC values, where the smallest AIC and BIC values are the best models for estimating the number of earthquakes. The results of the parameter estimation calculation process ($\boldsymbol{\lambda}$, $\boldsymbol{\delta}$, and $\boldsymbol{\Gamma}$) with hidden states $m = (2,3,4)$ can be seen in Table 7.

Table 7. EM algorithm estimation parameters for each PHMMs

m	- llk	AIC	BIC	i	$\hat{\lambda}$	$\hat{\delta}$	$\hat{\Gamma}$			
							1	2	3	4
m=2	155.12	320.24	330.00	1	5.67	1	0.84	0.16	-	-
				2	15.60	0	0.06	0.94	-	-
m=3	140.45	**302.91**	**324.38**	1	2.80	1	0.86	0.14	0.00	-
				2	9.66	0	0.00	0.95	0.05	-
				3	16.75	0				
m=4	140.20	318.40	355.46	1	2.79	1	0.86	0.14	0.00	0.00
				2	9.64	0	0.00	0.95	0.05	0.00
				3	15.70	0	0.00	0.00	0.70	0.30
				4	20.22	0	0.00	0.00	1.00	0.00

Based on Table 7, the estimated values using the EM algorithm in PHMMs, it can be seen that the smallest AIC and BIC values (marked with bold numbers) are in three hidden states, namely with AIC value = 302, 91 and BIC value = 324.38. So, it can be said that the 3 hidden states model is the best compared to the m=2 and m=4 models. After the model building phase was completed and the AIC and BIC values were determined for each model, PMMs, and PHMMs, both models resulted in the identification of the same number of unobserved groups, namely 3 groups. This can be seen by looking at the smallest AIC and BIC values of each model. Furthermore, for the case of many triplets, it was found that the PHMMs model had the smallest AIC and BIC values compared to PMMs. Therefore, the model chosen for the seismic data modeled in this study follows the probability distribution of PHMMs with a number of unobserved groups of 3. The following formulation of the density probability function of PHMMs for 3 hidden states is as follows:

$$p(x) = \sum_{i=1}^{m} \delta_i Poi(\lambda_i)$$

$$p(X_t) = \big((0.44)Poi(6.00) + (0.40)Poi(14.24) + (0.15)Poi(19.25)\big)$$

with the transition probability matrix of the unobservable data process is

$$\boldsymbol{\Gamma} = \begin{bmatrix} 0.74 & 0.17 & 0.09 \\ 0.24 & 0.57 & 0.19 \\ 0.00 & 0.71 & 0.29 \end{bmatrix}.$$

D. CONCLUSION AND SUGGESTIONS

We successfully applied two types of Poisson distribution (Poi (λ)), mixed models, namely the independent Poisson mixed model and the dependent Poisson model, to build a probabilistic model that uses historical data on the frequency of earthquake events from January 1st 1971, to December 31, 2022, in the Enggano segment of the Sumatra subduction zone. The estimation technique for the implemented probability model parameters uses the expectation and maximization algorithm, while the data used is data on the frequency of occurrence of major earthquakes with magnitude $M_w \geq 4.40$ obtained from the United States Geological Survey (USGS).

The modeled empirical data has an average value of $M_w = 12.17$ with a variance of $M_w = 38.18$. In other words, there is an overdispersion condition in the research data because the variance value is larger than the sample average. Based on the results of testing the dependence of research data on time using the Ljung-Box test, in which the *p-value* ($2.02\text{x}10^{-5}$) of the test statistic from the Ljung-Box test is less than the selected α -Significance is level value, namely 0.05, then evidence was found of the dependence of the frequency of earthquakes each year on time. From these two test results, it can be concluded that the probability model that fits the modeled data is a dependent mixed model of Poisson distribution, namely m-state Poisson Hidden Markov Models (PHMMs). To determine the number of states (m) from the m-state PHMMs model, we implemented a trial-and-error technique for multiple values m=2,3 and 4. Based on the model selection criteria using the smallest AIC and BIC With these values, we obtained the number of unobserved states or groups. The modeled data is 3 states, namely $p(x) = (044)\ Poi(6.00) + (0.40)\ Poi(14.24) + (0.15)\ Poi(19.25)$, with the transition probability matrix of the unobservable process is as follows:

$$\Gamma = \begin{bmatrix} 0.74 & 0.17 & 0.09 \\ 0.24 & 0.57 & 0.19 \\ 0.00 & 0.71 & 0.29 \end{bmatrix}.$$

In most coastal areas of Indonesia, there is a need for study results in the field of disaster risk reduction, especially modeling of earthquake events. Therefore, it is hoped that the results of this research can provide an additional contribution to complement the results of previous studies conducted by earthquake researchers in the Enggano segment to map the degree of earthquake vulnerability in this segment. Specifically, this article describes of unobserved behavior (dynamics of tectonic plates) based on selected models, however this study is still being carried out on a very small scale by earthquake researchers.

ACKNOWLEDGEMENT

The authors acknowledge the funding provided by the Faculty of Mathematics and Natural Sciences (FMIPA), The University of Bengkulu under the scheme of *Penelitian Unggulan* (Grant numbers: 1990/UN30.12/HK/2023).

REFERENCES

Boden, D. R. (2016). *Geologic fundamentals of geothermal energy*. CRC Press.

Daniell, J. E., Schaefer, A. M., & Wenzel, F. (2017). Losses associated with secondary effects in earthquakes. *Frontiers in Built Environment, 3*, 30. https://doi.org/10.3389/fbuil.2017.00030

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Ding, J., Tarokh, V., & Yang, Y. (2017). Bridging AIC and BIC: A new criterion for autoregression. *IEEE Transactions on Information Theory, 64*(6), 4024–4043. https://doi.org/10.1109/TIT.2017.2785860

George, S., & Jose, A. (2020). Generalized Poisson hidden Markov model for overdispersed or underdispersed count data. *Revista Colombiana de Estadística, 43*(1), 71–82.

Greff, K., Van Steenkiste, S., & Schmidhuber, J. (2017). Neural expectation maximization. In *Advances in neural information processing systems* (Vol. 30).

Hafiez, H. A. (2015). Estimating the magnitude of completeness for assessing the quality of earthquake catalogue of the ENSN, Egypt. *Arabian Journal of Geosciences, 8*, 9315–9323. https://doi.org/10.1007/s12517-015-1887-5

Hassani, H., & Yeganegi, M. R. (2020). Selecting optimal lag order in Ljung–Box test. *Physica A: Statistical Mechanics and Its Applications, 541*, 123700. https://doi.org/10.1016/j.physa.2019.123700

Irsyam, M., Cummins, P. R., Asrurifak, M., Faizal, L., Natawidjaja, D. H., Widiyantoro, S., Meilano, I., Triyoso, W., Rudiyanto, A., Hidayati, S., Ridwan, M., Hanifa, N. R., & Syahbana, A. J. (2020). Development of the 2017 national seismic hazard maps of Indonesia. *Earthquake Spectra, 36*(1), 112–136. https://doi.org/10.1177/8755293020951206

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research, 33*(2), 188–229. https://doi.org/10.1177/0049124103262065

Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika, 65*(2), 297–303. https://doi.org/10.1093/biomet/65.2.297

Mannering, F. L., Shankar, V., & Bhat, C. R. (2016). Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research, 11*, 1–16. https://doi.org/10.1016/j.amar.2016.04.001

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and Its Application, 6*, 355–378. https://doi.org/10.1146/annurev-statistics-031017-100325

Orfanogiannaki, K., Karlis, D., & Papadopoulos, G. A. (2014). Identification of temporal patterns in the seismicity of Sumatra using Poisson hidden Markov models. *Research in Geophysics, 4*(1), 1–6. https://doi.org/10.4081/rg.2014.4969

Radziminovich, N. A., Miroshnichenko, A. I., & Zuev, F. L. (2019). Magnitude of completeness, b-value, and spatial correlation dimension of earthquakes in the South Baikal Basin, Baikal Rift System. *Tectonophysics, 759*, 44–57. https://doi.org/10.1016/j.tecto.2019.03.011

Reid, A. (2015). History and seismology in the Ring of Fire: Punctuating the Indonesian past. In *Environment, trade and society in Southeast Asia* (pp. 1–16). Brill. https://doi.org/10.1163/9789004288058_006

Rizal, J., Gunawan, A. Y., Indratno, S. W., & Meilano, I. (2018). Identifying dynamic changes in megathrust segmentation via Poisson mixture model. In *Journal of Physics: Conference Series* (Vol. 1097, No. 1, Article 012083). IOP Publishing. https://doi.org/10.1088/1742-6596/1097/1/012083

Rizal, J., Gunawan, A. Y., Indratno, S. W., & Meilano, I. (2021). The application of copula continuous extension technique for bivariate discrete data: A case study on dependence modeling of seismicity data. *Mathematical Modelling of Engineering Problems, 8*(5), 793–804. https://doi.org/10.18280/mmep.080516

Rizal, J., Gunawan, A. Y., Indratno, S. W., & Meilano, I. (2023). Seismic activity analysis of five major earthquake source segments in the Sumatra megathrust zone: Each segment and two adjacent segments points of view. *Bulletin of the New Zealand Society for Earthquake Engineering, 56*(2), 55–70. https://doi.org/10.5459/bnzsee.1555

Sebastian, T., Jeyaseelan, V., Jeyaseelan, L., Anandan, S., George, S., & Bangdiwala, S. I. (2019). Decoding and modelling of time series count data using Poisson hidden Markov model and Markov ordinal

logistic regression models. *Statistical Methods in Medical Research, 28*(5), 1552–1563. https://doi.org/10.1177/0962280218761154

Wang, J., Zhao, J., Lei, X., & Wang, H. (2018). New approach for point pollution source identification in rivers based on the backward probability method. *Environmental Pollution, 241*, 759–774. https://doi.org/10.1016/j.envpol.2018.06.017

Williams, Q. (2018). The thermal conductivity of Earth's core: A key geophysical parameter's constraints and uncertainties. *Annual Review of Earth and Planetary Sciences, 46*, 47–66. https://doi.org/10.1146/annurev-earth-082517-010208

Yip, C. F., Ng, W. L., & Yau, C. Y. (2017). A hidden Markov model for earthquake prediction. *Stochastic Environmental Research and Risk Assessment*, 1–20. https://doi.org/10.1007/s00477-017-1443-1

Zucchini, W., MacDonald, I. L., & Langrock, R. (2017). *Hidden Markov models for time series: An introduction using R*. Chapman and Hall/CRC. https://doi.org/10.1201/b20790