Bulletin of the New Zealand Society for Earthquake Engineering, Vol. 56, No. 2, June 2023

SEISMIC ACTIVITY ANALYSIS OF FIVE MAJOR EARTHQUAKE SOURCE SEGMENTS IN THE SUMATRA MEGATHRUST ZONE: EACH SEGMENT AND TWO ADJACENT SEGMENTS POINTS OF VIEW

Jose Rizal[1], Agus Y. Gunawan[2], Sapto W. Indratno[3] and Irwan Meilano[4]

(Submitted *July 2021*; Reviewed *November 2021*; Accepted *December 2022*)

ABSTRACT

The Sumatra megathrust zone has five major earthquake sources, namely the Aceh-Andaman, Nias-Simeulue, Mentawai-Siberut, Mentawai-Pagai, and Enggano segments. This paper provides seismic activity analysis in these five segments via an unobserved process study of tectonic plate movements, which is conducted in two cases: each of the five segments independently (Case 1), and a pair of two adjacent segments (Case 2). To do this, two specific types of Hidden Markov Models (HMMs), i.e., Poisson-HMMs and Copula-HMMs, dealing with unobserved process issues, are applied. In practice, the data used is the annual frequency of mainshock earthquakes with a magnitude of $M_w \geq 4.6$ that occurred from 1971 to 2018. This data is obtained by working out the declustering process and estimating the magnitude of completeness from a particular earthquake catalogue. Due to the incompleteness of the data sets, the parameters of the two HMMs are estimated using the Expectation-Maximization algorithm. Results show that for Case 1, the model that fits the data for each of the five segments is the 3-state Poisson-HMM. The three states, in this instance, stand for the rates of seismic activity that correspond to the dynamic level of tectonic plate movements. Furthermore, in Case 2, the selected model for the Aceh-Andaman with Nias-Simeulue is the 2-state Gumbel Copula-HMM. Meanwhile, for the three groups remaining, namely Nias-Simeulue with Mentawai-Siberut, Mentawai-Siberut with Mentawai-Pagai, and Mentawai-Pagai with Enggano, the appropriate models are Gaussian, Gumbel, and Frank Copulas, respectively. In this case, the number of states represents the seismic activity association of two adjacent segments that corresponds to the association level of two adjacent tectonic plate dynamics.

https://doi.org/10.5459/bnzsee.1555

INTRODUCTION

In Indonesia, earthquakes are one of the biggest threats to life safety and the sustainability of public infrastructure due to the fact that this region is located on the Pacific Ring of Fire, which is a region with a high level of tectonic plate movements [1]. There are three active tectonic plates in Indonesia, namely the Eurasian, the Pacific, and the Indo-Australian Plates, that are still moving dynamically and sliding past each other. The meeting of these tectonic plates forms the subduction zone: an area with the potential for major earthquake occurrences [1,2]. One of these is the Sumatra megathrust zone, which contains five major earthquake sources, namely Aceh-Andaman (AA), Nias-Simeulue (NS), Mentawai-Siberut (MS), Mentawai-Pagai (MP), and Enggano (EO) segments, as shown in Figure 1 (a). The tectonic plate framework of these five segments has varied characteristics, where the area with the highest seismic activity is the AA segment and the lowest one is the EO segment [3]. For this reason, the AA segment is the most hazardous area in the Sumatran megathrust zone [2]. A detailed description of the tectonic frameworks and seismic hazards of the five major earthquake source segments in the Sumatra megathrust zone can be seen in Irsyam et al. [2] and McCaffrey [3].

Historically, the Sumatra megathrust zone has encountered major earthquakes frequently during the past two decades [4]. Some of them have caused a large number of deaths and serious infrastructure damage [4,5]. For example, the Sumatra-Andaman earthquake in the AA segment (December 26, 2004; M_w 9.1), the Nias-Simeulue earthquake in the NS segment (March 28, 2005; M_w 8.6), and the Padang earthquake in the MS segment (September 30, 2009; M_w 7.5) [4-6]. Furthermore, after two major earthquakes occurred in the AA and NS segments, the next one is predicted to occur around the MS and MP segments with magnitude $M_w > 8$, as mentioned by Sieh [7] and McCloskey et al. [8]. According to this information, a comprehensive study of earthquake risk management in the Sumatra megathrust zone through a seismic activity study has become an essential issue to be carried out [2].

Some previous earthquake researchers have discussed the issue from geological [9,10], seismological [11,12], and geophysical [13,14] points of view, just to name a few. Note that, however, this current study takes a different perspective to respond to the issue. That is, we provide a seismic activity study based on the analysis of an unobserved process of tectonic plate movements. The analysis is conducted in two cases, namely each of the five segments independently (Case 1), and a pair of two adjacent segments (Case 2). To accomplish this, Cases 1 and 2 are subjected to two types of Hidden Markov Models (HMMs) dealing with an unobserved process analysis, namely Poisson-HMMs [15-18] and Copula-HMMs [19,20]. The background that leads us to use these approaches is expressed as follows.

[1] *Corresponding Author, Associate Professor, Mathematics Department, Faculty of Mathematics and Natural Sciences, University of Bengkulu, Bengkulu, Indonesia, jrizal04@unib.ac.id.*

[2] *Associate Professor, Industrial and Financial Mathematics Research Group, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Bandung, Indonesia, ayodi@itb.ac.id.*

[3] *Associate Professor, Statistics Research Group, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Bandung, Indonesia, saptowi@itb.ac.id.*

[4] *Associate Professor, Faculty of Earth Sciences and Technology, Institut Teknologi Bandung, Bandung, Indonesia, irwanm@itb.ac.id.*

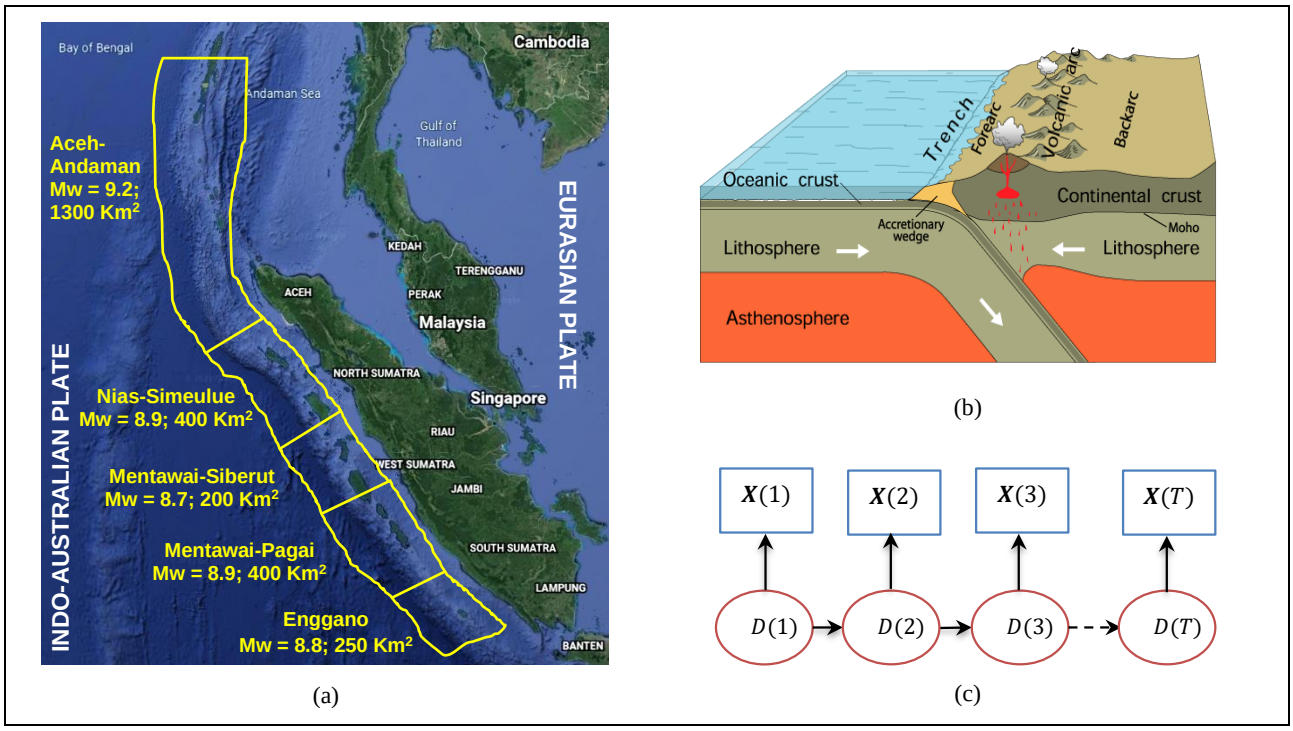

Figure 1: Part (a): a map of five segments of major earthquake sources in the Sumatra megathrust zone, with an estimated maximum earthquake, M_{max} (M_w), and area (km^2) [2]. Part (b): an illustration of the subduction process at plate boundaries (https://www.usgs.gov/). Part (c): a directed graph of the basic hidden Markov models (HMMs) [18].

Figure 1 (a) shows that earthquakes in the Sumatra megathrust zone depend on the tectonic plate movement mechanism between the Indo-Australian and Eurasian Plates that meet at the boundaries on the seabed [2]. Specifically, the Indo-Australian Plate is slowly subducting under the continental Eurasian Plate in the lithosphere due to convection currents in the asthenosphere, as shown in Figure 1 (b). Let us consider that $D(t) = \{D(t): t \in \mathrm{N}\}$ is the behavior of unobserved tectonic plate dynamics, which describes the levels of tectonic plate movements in finite categories, whereas $\boldsymbol{X}_i(t) = \{X_1(t), \dots, X_5(t): t \in \mathrm{N}\}$ is the random vector of earthquake frequencies from the five segments in the Sumatra megathrust zone. With this consideration, the dependence model between tectonic plate movements and earthquake frequencies can be represented in the form of a directed graphical model of $\{D(t), \boldsymbol{X}_i(t)\}$, as illustrated in Figure 1 (c). However, note that there are drawbacks in the implementation, namely that while earthquakes were observed, the dynamic of tectonic plates was not observed directly [15-17]. This condition makes the mathematical modeling of the earthquake issue challenging, especially in seismic activity studies. One of the methods that can be used to solve this problem is the hidden Markov models (HMMs). This model is powerful for modeling generative unobservable sequences that can be characterized by an underlying process generating an observable sequence [18].

Furthermore, if we assume that the variable $X_i(t)$ is generated from a Poisson distribution and depends only on the current state $D(t)$ (Case 1), with m states, then the seismic activity of each segment can be analyzed independently using the m-state Poisson-HMMs [15-17]. The number of states m in this case represents the seismic activity rates that correspond to the number of levels of dynamic tectonic plate movement (e.g., low, medium, or high dynamics) [2,3]. Furthermore, if the condition of spatial dependence between two adjacent segments exists (case 2), then to resolve this issue, the Copula models can be applied [19,20]. Therefore, the earthquake issue in setting consideration for case 2 is performed by a mixture of Copula models and HMMs, denoted by the m-state Copula-HMMs. Here, the number of states m represents the seismic activity associations of two adjacent segments and corresponds to the level of association of two adjacent tectonic dynamics, e.g., weak, moderate, or strong associations. The parameters of the two probability models are estimated based on the seismicity data, which is obtained from a particular earthquake catalog.

The seismicity data commonly used in seismic activity studies is the earthquake frequencies with a certain magnitude threshold and time interval [15-17,21,22]. In other words, the characteristics of the seismicity data are categorized as discrete variables. However, this condition becomes a serious problem in performing the second case due to the non-uniqueness of the Copula models for marginal discrete variables [23]. To overcome this problem, the marginal variables should be transformed into continuous variables using the Continuous Extension technique developed by Denuit and Lambert [24]. Thus, the Copulas parameter estimation of continued variables can be worked out using the same procedure as in the continuous variables issue [25]. The advantage of this technique is used in the current study.

To the best of our knowledge, the probability models that we proposed, in particular the m-state Copula-HMMs that worked with the Continuous Extension technique, are rarely discussed in earthquake engineering studies. Thus, it is hoped that the results of this study will make a significant contribution to the earthquake engineering literature and the earthquake risk management issues, especially in the Sumatra megathrust zone.

The outline of this paper is divided into five sections. Section 1 gives an introduction to the seismic issues in the Sumatra subduction zone and the background of using the m-state Poisson-HMMs and m-state Copula-HMMs approaches. Section 2 provides a brief introduction to the mathematical tools, namely HMMs, Poisson distribution, Copula models, and Continuous Extension technique, which will be used in this paper in a self-contained manner. In Section 3, we describe the seismicity data preparation and the data analysis procedure. Meanwhile, the results and discussion are presented in Section 4. Finally, conclusions are written in Section 5.

MATHEMATICAL TOOLS DESCRIPTION

Four mathematical concepts, namely the hidden Markov models, Poisson distribution, bivariate Copula models, and the Continuous Extension technique, are discussed in this section. Furthermore, a detailed discussion of those concepts as well as a description of their properties can be seen in the books and articles that we refer to. For the first and second concepts, we refer to the book by Zucchini et al. [18]. For the third concept, we refer to the books of Joe [19], Nelsen [20], and Hofert et al. [26]. For the last concept, we refer to the articles of Denuit and Lambert [25], Stevens [27], and Rizal et al. [28].

Hidden Markov Models

From a theoretical perspective, HMMs are defined as discrete-time processes that consist of two random processes [18]. The first part is a finite-state Markov chain $D(t)$ on m states, which is unobservable and thus referred to as a hidden process. In addition, the sequence $D(t)$ satisfies the Markov property; i.e., given the present, the future does not depend on the past. The second part is an observable of the state-dependent process $X(t)$ such that, for a given $D(t)$, the distribution of $X(t)$ depends only on the present state $D(t)$ and not on previous or observations. This schema is represented graphically in part (c) of Figure 1, which can be applied to univariate and multivariate cases.

The parameters of HMMs are characterized in three parts, namely the transition probability matrix of the Markov chain, stationary distribution, and probability distribution that are related to the states $D(t)$. Each state is associated with a probability density function f from the same parametric family. When $D(t) = d(t)$ is given, $X(t)$ takes the values $x(t)$ with probability $P(x(t)|d(t))$. In the univariate case, for instance, if we assume that each observation is generated from a discrete distribution, namely the Poisson distribution, then probability takes the form $P(x(t)|d(t) = s)$ = Poisson (λ_s) where $\lambda_s \geq 0$ are the parameters of the Poisson distribution and $x(t)$ is a non-negative integer value, for all $t = 1,2, \dots, T$ and $s = 1,2, \dots, m$. Meanwhile, for the bivariate case $\boldsymbol{X} = \big(X_1(t), X_2(t)\big)$, if it is assumed that $\boldsymbol{x} = \big(x_1(t), x_2(t)\big) \in R^2$ is generated from a joint distribution of two variables $g(x_1, x_2;\ \vartheta)$, with ϑ represents the set parameter model, then f takes the form:

$$P(\boldsymbol{x}(t)|d(t) = s) = g(x_1, x_2;\ \vartheta_s), \tag{1}$$

In this article, we only determine the likelihood function of HMMs in the bivariate case, as calculated under m-state HMMs, which has the transition probability matrix $\boldsymbol{\Gamma} = \gamma_{rs} = P(D(t) = s|D(t-1) = r)$, stationary distribution $\boldsymbol{\delta}$ implied by $\boldsymbol{\Gamma}$, and state-dependent probability function $g_{\vartheta_{c(t)}} = g(x_1, x_2;\ \vartheta_s)$. Subsequently, if we denote with $\boldsymbol{\Psi} = (\vartheta_1, \vartheta_2, \dots, \vartheta_m, \boldsymbol{\Gamma})$ the parameter of models to be estimated, then the complete-data log-likelihood of HMMs is as follows:

$$\log\big(P(\boldsymbol{X}, D)\big) = \log\left(\boldsymbol{\delta}_{c(1)} \prod_{t=2}^{T} \gamma_{d(t-1),d(t)} \prod_{t=1}^{T} g_{\vartheta_{c(t)}}(\boldsymbol{x}(t))\right) \tag{2}$$

In the HMMs, a set of states and the transition probability matrix are treated as missing data. Therefore, we first need to define two random variables, $u_j(t)$ and $v_{jk}(t)$, that correspond to the sequence states $d(1), d(2), \cdots, d(T)$ by the zero-one random variables. The random variable $u_j(t)$ is equal to 1 if the state of the model at time t is j and 0 otherwise, $t = 1, 2, \dots, \mathrm{T}$. Whereas the random variable $v_{jk}(t)$ is equal to 1 if $d(t-1) = j$ and $d(t) = k$, and 0 otherwise, with $t = 2,3, \dots, \mathrm{T}$. Using those two random variables, namely, $u_j(t)$ and $v_{jk}(t)$, we get the complete log-likelihood formulation of the HMMs from Equation (2), as follows:

$$\begin{aligned}\log\big(P(\boldsymbol{X}, D)\big) &= \sum_{j=1}^{m} u_j(1) \log \delta_j + \sum_{j=1}^{m}\sum_{k=1}^{m}\left(\sum_{t=2}^{\mathrm{T}} v_{jk}(t)\right) \log \gamma_{jk} \\ &\quad + \sum_{j=1}^{m}\sum_{t=1}^{\mathrm{T}} u_j(t) \log g_{\vartheta_{c(t)}}(\boldsymbol{x}(t)).\end{aligned} \tag{3}$$

Since the states $D(t)$ in the HMMs are treated as missing data (unobservable), one way to perform parameter estimates of Eq. 3 is to use the Expectation-Maximization (EM) algorithm that was rigorously derived in [29]. This proceeds in two steps, namely the E step, where $\mathcal{H}_{\boldsymbol{x}}\big(\bar{\boldsymbol{\Psi}}, \boldsymbol{\Psi}\big) = E_{\boldsymbol{\Psi}}[\log P(\boldsymbol{X}, D)|\boldsymbol{X} = \boldsymbol{x}]$ is calculated, and the M step, where one is calculated as follows:

$$\boldsymbol{\Psi}^{(l+1)} = \arg\max_{\boldsymbol{\Psi}} \mathcal{H}_{\boldsymbol{x}}\big(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(l)}\big), \tag{4}$$

with l expressing the iteration and starting from an initial value $\boldsymbol{\Psi}^{(0)}$. While $l \to \infty$, $\boldsymbol{\Psi}^{(l)}$ converges to the maximum likelihood estimator of the density of $\boldsymbol{X}$. The detailed steps of the EM algorithm are described in Nasri et al. [30] and are omitted to save space. In practice, the parameters estimation of Copula-HMMs is worked out using the R package "HMMcopula" developed by Thioub et al. [31]. Meanwhile, for the Poisson HMMs case, the parameter estimation is carried out using R codes of "Pois-HMM" that are provided by Zucchini et al. [18].

It should be highlighted that determining the optimal state sequence, which would most adequately describe the observation sequence in some way, is another challenge that needs to be resolved in the context of HMMs [18]. That is, given as input the parameters of HMMs $\boldsymbol{\Psi}$ and a sequence $\{X(t)\}$ of length T, we want to find the most likely sequence of states $\{D(t)\}$ that produces the output sequence $\{X(t)\}$. This means finding a sequence $\{D(t)\}$ such that the probability of Eq. 2 is maximal. This problem is solved effectively by the Viterbi algorithm [32], and the details are as follows.

To convert the products form in Eq. 2 into summations form, we define $A(D(t))$ first as

$$A(D(t)) = -\log\left(\boldsymbol{\delta}_{d(1)} g_{\vartheta_{d(1)}}(\boldsymbol{x}(1))\right) - \left[\sum_{t=2}^{T} log\ \left(\gamma_{d(t-1),d(t)}\ g_{\vartheta_{d(t)}}(\boldsymbol{x}(t))\right)\right]. \tag{5}$$

Consequently, the optimal state estimation problem for $\max_{D(t)} P(\boldsymbol{X}, D|\boldsymbol{\Psi})$ becomes equivalent to $\min_{D(t)} A(D(t))$ [32]. Subsequently, let $A_t(D(1), D(2), \dots, D(t))$ be the first t terms of $A(D(t))$ and $B_t(i)$, which denotes the minimal accumulated cost when we are in state i at time t, namely

$$B_t(i) = \min_{D(1),\dots,D(t)=i} A_t(D(1), D(2), \dots, D(t) = i). \tag{6}$$

Viterbi algorithm then can be implemented in four steps as follows:

1. Initialize the $B_t(i)$ for all $1 \leq i \leq m$:
$$B_t(i) = -\log\left(\boldsymbol{\delta}_{d(1)} g_{\vartheta_{d(1)}}(\boldsymbol{x}(1))\right)$$
2. Inductively calculate the $B_t(i)$ for all $1 \leq i \leq m$, from time $t = 2$ until $t = T$:
$$B_t(i) = \min_{1 \leq j \leq m} B_{t-1}(j) - \log\left(\gamma_{d(t-1)=j,d(t)=i}\ g_{\vartheta_{d(t)}}(\boldsymbol{x}(t))\right)$$
3. Termilize the minimal value of $A(D(t))$:
$$\min_{D(t)} A(D(t)) = \min_{1 \leq i \leq m} B_T(i)$$
4. Finally, we trace back the calculation to find the optimal state of $D(t)$.

From the aspect of state-dependent probability, we then notice that the Copulas were used because of their flexibility to determine the joint probability density function $g_{\vartheta_{d(t)}}$ [19,20]. Therefore, a more thorough explanation of the Copulas theory is required, which will be provided in the following subsection.

Bivariate Copula Models

As is well known, the Copula models or Copulas (denoted by C) are one of the tools in dependency modeling that are quite popular in a wide range of scientific fields, such as engineering [33], economics [34], geostatistics [35,36], and finance [37,38]. While Copulas are complicated to spell out in multivariate variables, it becomes relatively easy to discuss in the bivariate case [20]. As previously mentioned in Section 1, this study is focused on modeling the dependence of two adjacent segments. For these reasons, the discussion related to Copulas is restricted to the bivariate Copula, which is defined as a two-dimensional cumulative distribution function (CDF) of variables wherein dependence between marginals exists.

Generally, the joint distribution of two random variables X_1 and X_2 with CDFs of $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$ is denoted by $F_{X_1,X_2}(x_1,x_2)$. Meanwhile, the bivariate Copula is denoted by $C\big(F_{X_1}(x_1), F_{X_2}(x_2);\ \theta\big)$, where the notation of θ is the Copula parameter that describes a measure of dependency. By following Sklar's theorem [39], the relationship of both joint distributions can be expressed as follows:

$$F_{\mathrm{X}_1,\mathrm{X}_2}(x_1,x_2) = C\big(F_{\mathrm{X}_1},F_{\mathrm{X}_2};\theta\big) = C(u_1,u_2;\theta). \qquad (7)$$

The theorem of integral transform probability [40] shows that the marginal CDF of variables X_1 and X_2 follows the Uniform distribution ranging from 0 to 1, denoted by $u(0,1)$. Therefore, $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$ can be rewritten as u_1 and u_2, respectively.

In many kinds of Copulas literature, a number of bivariate Copula functions have been widely used in practice. However, the results in this paper rely on the following four commonly used, namely Gaussian, Clayton, Gumbel, and Frank Copulas. Table 1 shows their summary and mathematical expression. In addition, the Copula parameter is estimated using a two-stage procedure called Inference Function for Marginals in which the marginal distribution is estimated in the first step and then we treat them to estimate the parameter Copula θ, in either a full or semi-parametric approach [41].

Table 1: Summary of the mathematical formulation of Copula model, and domain parameter for Gaussian, Clayton, Gumbel, and Frank copulas.

Copula Models	$C(u_1,u_2;\theta)$	Domain of θ
Gaussian	$\Phi_G[\Phi^{-1}(u_1),\Phi^{-1}(u_2);\theta]$	$(-1,1)$
Clayton	$\left(u_1^{-\theta}+u_2^{-\theta}-1\right)^{-1/\theta}$	$(0,\infty)$
Gumbel	$\exp\left(-\left((-\log u_1)^{\theta} + (-\log u_2)^{\theta}\right)^{1/\theta}\right)$	$[1,\infty)$
Frank	$-\frac{1}{\theta}\log\left(1+\frac{e^{(-\theta u_1-1)}\,e^{(-\theta u_2-1)}}{\exp(-\theta)-1}\right)$	$(-\infty,\infty)$ \{0}

Notes: The notation of Φ is the cdf of $N(0,1)$, meanwhile Φ^{-1} is functional inverse of Φ. The symbol of $\Phi_G[*,*;\theta]$ is the bivariate normal distribution with correlation θ.

For the semi-parametric approach, the estimated margins from bivariate random variables were typically used to form the sample $\boldsymbol{u}_{1:T,j} = \left(F_{1:T,1}(X_1(t)), F_{1:T,2}(X_2(t))\right)$. It is known as pseudo-observation, and it is formulated as follows:

$$F_{1:T,j}(x) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{1}\big(X_j(t)\le x\big),\ x\in\mathbb{R},\ j=1,2 \qquad (8)$$

where T represents the observation length, $\mathbf{1}$ is the indicator function, that is, $\mathbf{1}\big(X_j(t)\le x\big)$ is one when $X_j(t)\le x$ and zero otherwise. This approach can be avoided in relation to any misspecification of the estimated margins [42].

Note that the theory of Copulas as outlined above is applied to continuous variables. In the case of discrete variables, it is not expressed as Eq. 7 due to a lack of uniqueness [23,39]. Also, there is another serious problem related to the estimate of the Copulas parameter θ. That is, the marginal distribution jumps when either F_{X_1} or F_{X_2} is noncontinuous. It causes the inverses ${F_{X_1}}^{-1}$ and ${F_{X_2}}^{-1}$ to have plateaus, which may lead to biased estimates of θ [43]. Therefore, these conditions have some disadvantages when applying the Copulas for discrete variables. One idea to overcome this problem is to analyze the continued discrete variables using the Continuous Extension technique [24]. With this idea, the estimation parameter Copulas of continued variables can be performed based on the procedure for continuous variables. This technique was also mentioned in Heinen and Rengifo [44], Nikoloulopoulos [45], and Inouye et al [46].

Continuous Extension Technique

This subsection introduces the concept of the Continuous Extension technique in transforming the discrete random variable X to a continuous, which is denoted by X^* [27]. Specifically, the former discrete variable becomes continuous by adding a random independent perturbation variable that is defined in [0,1]. This technique was used by Denuit and Lambert [25] in concordance measures for bivariate discrete data with some development. Additionally, the Uniform distribution (0,1) was selected as the most natural choice for a random independent perturbation, where it satisfies all the necessary constraints of the probability functions.

Here, we focus on the technique proposed by Denuit and Lambert [25]. They associate the discrete variable X with a continued variable X^* and a Uniform distribution U such that

$$X^* = X + (U-1). \qquad (9)$$

Variable U represents a Uniform distribution (0,1) and is independent of X. It is clear to see that this new random variable X^* is continuous and $X^*\le X$. Related to this current work, the essential result from Denuit and Lambert study is that the Continuous Extension technique preserves the dependence structure of the bivariate discrete variable [25,28].

THE SEISMICITY DATA PREPARATION AND THE DATA ANALYSIS PROCEDURE

The Seismicity Data Preparation

The seismicity data source is taken from the U.S. Geological Survey and Indonesia's Agency for Meteorology, Climatology, and Geophysics (BMKG) catalogs, including the foreshock, mainshock, and aftershock of earthquakes. However, by recalling the issue of the current study, only mainshock earthquakes are used in data analysis. Additionally, we need to determine a certain magnitude threshold (M_w) to work out the annual frequency tabulation process of mainshock earthquakes. According to these objectives, two stages of the seismicity data preparation will be performed, namely the declustering process [47] and the estimation value of the magnitude of completeness (M_c) [48]. In practice, both processes are performed using the Zmap package developed by Weimer [49].

Table 2: The standard input parameters for declustering algorithm by Reasenberg.

Notation	Parameter	Simulation range		Standard
	Represent	Minimum	Maximum	
τ_{min}	The minimum value (days) of the look-ahead time for building clusters when the first event is not clustered.	0.50	2.50	1.00
τ_{max}	The maximum value (days) of the look-ahead time for building clusters.	3.00	15.00	10.00
p_1	The probability detecting the next clustered event used to compute the look-ahead time.	0.90	0.99	0.95
x_k	The increse of the lower cut-off magnitude during clusters: $x_{meff} = x_{meff} + x_k M$, where M is the magnitude of the largest event in the cluster.	0.00	1.00	0.50
x_{meff}	The effective lower magnitude cutoff for catalog.	1.50	1.80	1.60
r_{fact}	The number of crack radii surrounding each earthquake within new events considered to be part of the cluster.	5.00	20.00	10.00

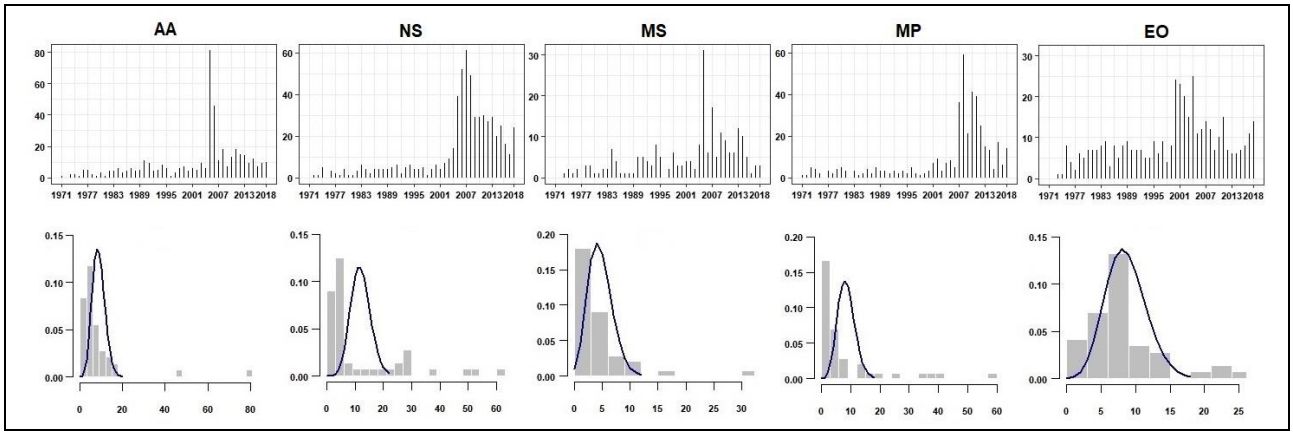

Figure 2: The visualizing of the annual frequencies of mainshock earthquakes with $M_w \geq 4.6$. Columns correspond to the five segments studied. Meanwhile, Rows correspond to bar plots (first row) and probability density function (second row).

Table 3: Descriptive statistics of the data.

Segments	Descriptive statistics					
	Min	Max	Mean	Var	Skew	Kurt
AA	0	81	8.73	165.12	4.32	23.33
NS	0	61	11.94	221.41	1.69	5.09
MS	0	31	4.58	27.56	3.04	14.93
MP	0	59	8.48	150.31	2.48	8.88
EO	0	25	8.58	32.04	1.24	4.59

Notes: Minimum (Min), Maximum (Max), Variance (Var), Skewness (Skew), and Kurtosis (Kurt).

For the first step, the Reasenberg algorithm that connects earthquakes according to the spatial and temporal distances among them [50], is applied. Technically, this algorithm eliminates earthquakes that occur within a specific distance and time, including foreshocks and aftershocks. There are six input parameters required for this, where the value of the six parameters in this study is determined by the standard value of each parameter (the last column of Table 2). Furthermore, for the second step, there are two common methods to estimate the M_c: the catalog-based [51] and network-based [52] methods.

Here, we concentrate on the former, that is, fitting the Gutenberg-Richter model to the observed frequency magnitude distribution, using this relation

$$\log_{10} N = a - b(m - M_c). \quad (10)$$

The used variables in Eq. 10 are as follows. Variable N is the number of earthquakes with a magnitude greater than or equal to m, a is the earthquake productivity, and b describes the relative distribution of small and large earthquake events.

The results of the seismic data preparation process using historical earthquake data from January 1971 to December 2018 are as follows: The estimated value of M_c is 4.6, and there were 419, 573, 220, 407, and 412 mainshock earthquakes for the AA, NS, MS, MP, and EO segments, respectively. The detailed data sets of the annual frequencies for the five segments are displayed graphically in Figure 2, while the descriptive statistics are provided in Table 3.

According to the first row of Figure 2, there are some periods within each segment with relatively low, moderate, and high rates. In addition, the condition of overdispersion relative to the Poisson distribution in the observed data exists for all segments, i.e., the variance is greater than the mean (see columns 4 and 5 of Table 3). From these conditions, it suggests that the annual earthquake frequencies of each segment may consist of unobserved groups that have a distinct distribution and are serial time-dependent [18] (result 1). Furthermore, since the skewness and kurtosis are positive and greater than 3 for all segments, the probability density empiric of each segment tends to be right-skewed and has fatter tails (result 2). These two results indicate the inappropriateness of the Poisson distribution with a single peak as a model. This is confirmed by the second row of Figure 2, which shows earthquake frequency data (bar chart) uncovered by the Poisson distribution's probability density function (pdf) with a single peak (dark blue color) due to unobserved heterogeneity in the data. For those reasons, we apply the HMMs to complete two cases of our research problem, which are dealing with unobserved heterogeneity and serial time-dependent issues in the data observations [18].

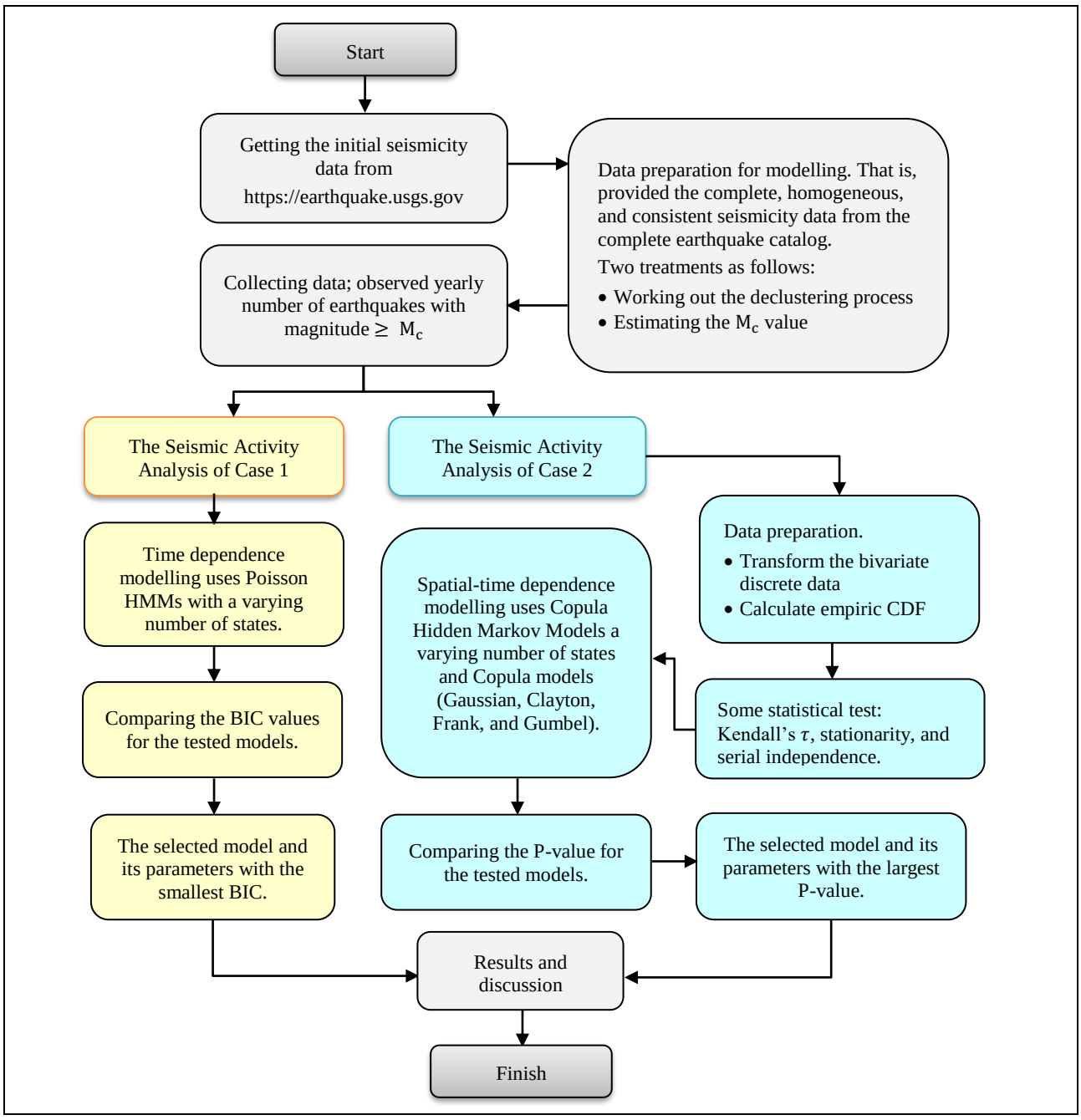

Figure 3: Flowchart of the data analysis procedure.

RESULTS AND DISCUSSION

In this section, we present the results and discussion of the seismic activity analysis of five major earthquake source segments in the Sumatra megathrust zone. In practice, we present it in two cases, namely each of the five segments independently (Case 1) and two adjacent segments (Case 2). For this purpose, two particular types of HMMs, namely Poisson-HMMs, which are used to complete Case 1, and Copula-HMMs, which are used to complete Case 2, are applied. However, to help the readers understand the steps of our work, we previously provided a flowchart diagram of the data analysis procedure, as shown in Figure 3.

The Seismic Activity Analysis of Case 1

Since the data modeled using the HMMs for Case 1 issue is assumed to have time-dependent properties, we first perform a serial time dependence test using the Ljung-Box test [53], where the detailed steps are seen in the Appendix of this article.

The null hypothesis (H_0) to be tested is that the annual frequency data is serially time-independent. It can be formulated formally as H_0: $\hat{\rho}^2(h) = 0$. The H_0 is rejected with a significance level of α, if the P-value of $Q_{LB}(h) < \alpha$. In this work, we choose $\alpha = 0.05$. The notation of $\hat{\rho}^2(h)$ is the sample autocorrelation at lag h, whereas $Q_{LB}(h)$ is the statistical test of the Ljung-Box test. The detailed formulas of those two notations are shown in Eq. A1 of the Appendix.

The results of the serial time dependence test for five segments are seen in Table 4, which shows that H_0 is rejected for each segment. Thus, the serial time dependence property of the data from each segment of five was present. This preliminary result is evidence that the Poisson-HMMs are capable models to be used in performing the issue of Case1 due to their ability to deal with time-series discrete data that is serially time-dependent. As a follow-up, the next step is to select the appropriate model of the Poisson-HMMs (i.e., determine the number of states m) and present its estimated parameters. We also processed the data with time-independent Poisson distributions such as Poisson

distribution (Pois.Dist), and Independent Poisson Mixture Models (Poisson-MMs) [18].

Table 4: Serial time dependence test of the annual mainshock earthquakes frequency using Ljung-Box Test.

Segments	Ljung-Box Test			
	$\hat{\rho}^2(1)$	$Q_{LB}(1)$	P-value	Decision of H_0
AA	0.441	9.994	1.61×10^{-03}	Rejected
NS	0.889	40.389	2.08×10^{-10}	Rejected
MS	0.308	4.867	2.74×10^{-02}	Rejected
MP	0.692	24.426	7.72×10^{-07}	Rejected
EO	0.635	20.593	5.68×10^{-06}	Rejected

Information-theoretical methods, such as Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC), can be used to select the appropriate model for any given data set when the model parameters are estimated using the maximum likelihood method. [18,54]. The following is their formula: $\mathrm{AIC}(m) = -2.ll(m) + 2l(m)$, and $\mathrm{BIC}(m) = -2.ll(m) + l(m).\ln(T)$, where $ll(m)$ is the value of log-likelihood, $l(m)$ is the number of parameters, and T is the size of observations.

Table 5 shows the AIC and BIC values for the three applied models, namely Poisson Distribution, Poisson-MMs, and Poisson-HMMs. Generally, for each segment analyzed, the smallest AIC(m) and BIC(m) belong to the Poisson-HMMs with the corresponding number of states (bold mark). Meanwhile, the Poisson distribution and Independent Poisson Mixture Models (Poisson-MMs) are not appropriate for the data modeled. What is noticeable from this result is that the selected model (Poisson-HMMs) for each of the five segments is consistent with the serial time dependence test; that is, the data modeled has serially time-dependent properties (see Table 4).

Subsequently, as can be seen in the 9th and 10th rows of Table 5, it is clear that for the MS and MP segments, the two model selection criteria (i.e., AIC and BIC) give the same results namely, 3-state Poisson-HMM. Meanwhile, for the AA, NS, and EO segments, it is seen that there is a different result between these two information criteria. As a response to this condition, we decided to use the BIC criteria in determining the final selected model. The background for choosing BIC to select the final model due to this criterion is a particular kind of model selection among a class of parametric models with various parameter numbers and is more conservative [18,54].

Table 5: Comparison of the fitted models by AIC and BIC.

The Models	m	Segments									
		AA		NS		MS		MP		EO	
		AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
Pois.Dist	-	638.077	639.949	878.304	880.175	344.923	346.794	714.985	716.857	352.275	354.147
Poisson-MMs	2	340.713	346.326	367.426	373.040	271.038	276.652	335.125	340.739	323.921	329.535
	3	306.248	315.604	332.445	341.801	257.354	266.710	294.560	303.915	289.416	298.772
	4	306.278	319.376	336.445	349.543	258.525	271.623	297.294	310.392	293.080	306.179
Poisson-HMMs	2	338.533	346.018	331.040	338.524	261.088	268.573	319.273	326.758	312.713	320.198
	3	280.022	**296.862**	287.573	**304.413**	**247.589**	**264.430**	**275.989**	**292.830**	268.530	**285.371**
	4	**277.508**	307.447	**285.271**	315.211	251.425	281.364	279.418	309.357	**268.083**	298.022

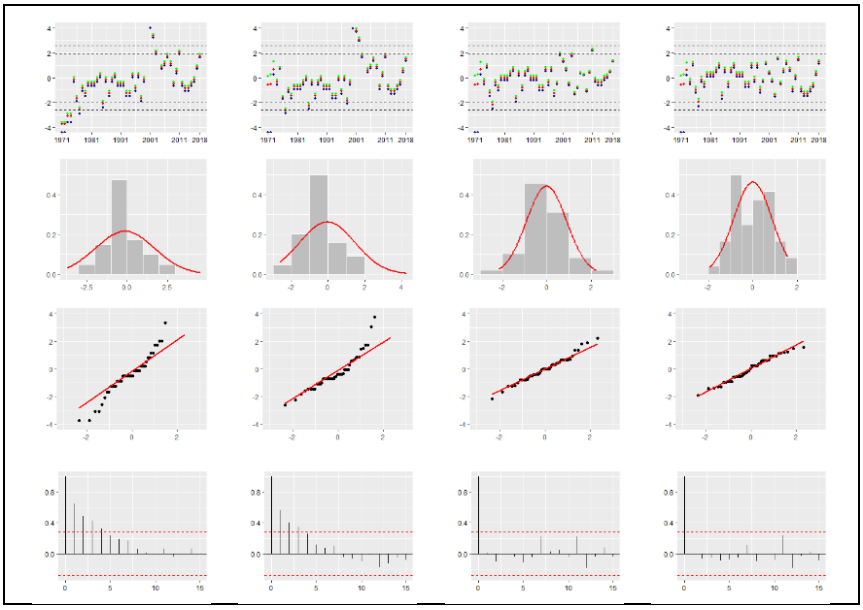

Figure 4: Diagnostic plots based on ordinary pseudo-residuals of the EO segment case. Rows correspond to (1) index plots of the normal pseudo-residuals with horizontal lines at $\pm$ 1.96 (2.5% and 97.5%) $\pm$ 2.58 (0.5% and 99.5%), (2) histograms of the normal pseudo-residuals with theoretical Normal distribution curves in red, (3) quantile-quantile plots (QQ-plots) of the normal pseudo-residuals with the theoretical quantiles on the horizontal axis, and (4) autocorrelation functions of the normal pseudo-residuals. Columns correspond to the Poisson-HMMs fitted to the EO segment data with 1,2,3, and 4 states respectively.

Furthermore, to check whether the best model provides an adequate fit to the data modeled, we analyze the normal pseudo-residuals of the selected model, including index plots, histograms, QQ-plots, and autocorrelation functions. That is, by comparing the normal pseudo-residuals of the 3-state Poisson-HMM as the selected model to those of the models with one, two, and four states. To do so, we apply the relevant R codes of the pseudo-residuals analysis for univariate data that was provided by Zucchini et al. [18]. To save space, we will focus the discussion only on the EO segment. Meanwhile, for other segments, it can be worked out with the same technique.

Let us consider the first and second columns of the ordinary pseudo-residuals of the EO segment displayed in Figure 4. Three objects can be discussed from these figures. The first is the presence of index plots elements of the normal pseudo-residuals that lie outside the 0.5% and 99.5% bands; the second is that the normal pseudo-residuals deviate significantly from the standard normal distribution; and the third is the presence of significant residual correlation on the normal pseudo-residuals for some lag. From these three conditions, it is clear that the single Poisson distribution and 2-state Poisson-HMM are not appropriate models for our data.

Additionally, for three and four-state Poisson-HMMs, three points can be stated from these models, such as: no observations stand out as extreme in the normal pseudo-residuals index plots; histograms and QQ-plots of the normal pseudo-residuals are approximately normal-shaped; and autocorrelations are near-zero, indicating low correlation in the residuals. By considering the model selection criteria and the diagnostic plots based on ordinary pseudo-residuals, it can be concluded that the 3-state Poisson-HMM is not only the selected model with the smallest BIC but also provides an adequate fit to the data.

We now present the estimated model parameters for each of the five segments fitted by the EM algorithm, which are summarized in Table 6. Table 6 shows the parameter estimates of the 3-state Poisson HMM, and its rows correspond to a different state for each. Specifically, the 3rd to 7th columns represent the parameters of the HMMs, namely the stationary distribution ($\boldsymbol{\delta}_P$), seismic rates ($\boldsymbol{\lambda}_P$) with their confidence intervals, and transition probability matrix (TPM) ($\boldsymbol{\Gamma}_P$). Also, the mean-variance of the observations and model are provided in the last four columns.

A bootstrap sample of size 200 was generated based on the 3-state Poisson HMM for the earthquake frequency data. The resulting sample of parameters then produced the 90% confidence intervals for the seismic rates parameter that are displayed in the 5th and 6th columns of Table 6. From the seismic rates parameter $\boldsymbol{\lambda}_P$ (i.e., the 4th through 6th columns), it emerges that for each segment, the intervals for the state-dependent mean $(\boldsymbol{\lambda}_P)_i$ do not overlap. Thus, it suggests that the states for each segment are well defined. In this work, the three states represent the seismicity rates of mainshock earthquake frequencies, namely "rare", "moderate", and "frequent" which correspond to the level of the tectonic movement, that is, "low", "medium", and "high" dynamics, respectively [2,3].

Table 6: The estimated parameters and means-variances of the selected models for the five segments.

Segments	State	Parameters					Observation		Model	
		$\boldsymbol{\delta}_P$	$\boldsymbol{\lambda}_P$			$\boldsymbol{\Gamma}_P$	Mean	Variance	Mean	Variance
			EM	90% Conf. Limits						
AA	1	0.620	4.383	3.638	5.164	$\begin{bmatrix} 0.982 & 0 & 0.018 \\ 0.032 & 0.968 & 0 \\ 0 & 0.506 & 0.494 \end{bmatrix}$	8.729	165.095	8.415	91.409
	2	0.358	11.918	9.182	14.227					
	3	0.022	63.500	54.104	71.661					
NS	1	0.606	3.567	2.959	4.485	$\begin{bmatrix} 0.981 & 0.019 & 0 \\ 0.038 & 0.875 & 0.087 \\ 0 & 0.270 & 0.730 \end{bmatrix}$	11.938	221.336	13.867	231.098
	2	0.298	23.085	21.215	24.850					
	3	0.096	50.339	43.965	55.648					
MS	1	0.776	2.534	2.084	2.998	$\begin{bmatrix} 0.957 & 0.043 & 0 \\ 0.175 & 0.655 & 0.170 \\ 0 & 1 & 0 \end{bmatrix}$	4.583	27.525	4.261	22.125
	2	0.192	7.842	6.459	9.063					
	3	0.032	24.317	19.913	29.200					
MP	1	0.579	2.621	1.542	4.022	$\begin{bmatrix} 0.980 & 0.020 & 0 \\ 0.040 & 0.881 & 0.079 \\ 0 & 0.185 & 0.815 \end{bmatrix}$	8.479	150.255	8.810	129.760
	2	0.295	9.015	7.422	10.551					
	3	0.126	36.839	33.703	39.396					
EO	1	0.214	0.498	0.000	1.269	$\begin{bmatrix} 0.933 & 0.067 & 0 \\ 0 & 0.936 & 0.064 \\ 0.063 & 0.094 & 0.843 \end{bmatrix}$	8.583	32.035	8.060	39.939
	2	0.558	7.199	5.947	8.470					
	3	0.228	17.235	15.322	20.098					

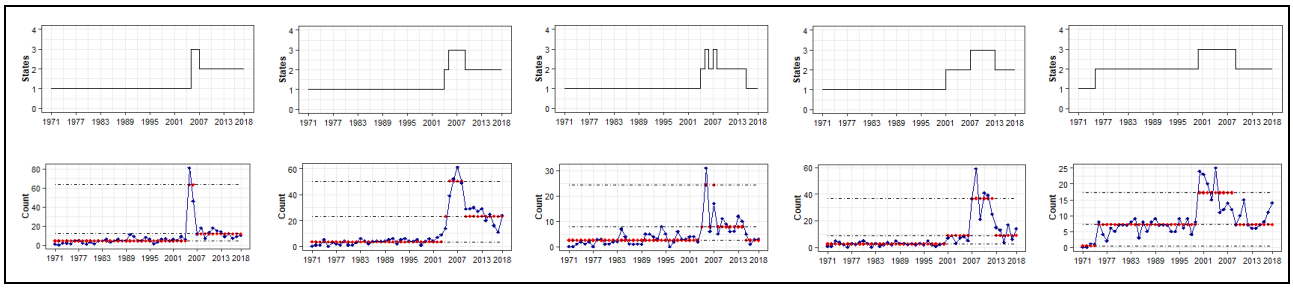

Figure 5: Rows correspond to (1) estimated states of seismic activity and (2) the observation compared with the state-dependent means of 3-state Poisson HMM. Columns correspond to the five segments, namely AA, NS, MS, MP, and EO respectively.

The visualization of the level of tectonic plate dynamic and the comparison between the observations as well as the mean of the selected models are displayed in Figure 5.

From Table 6 and Figure 5, some aspects can be discussed. For instance, let us consider the AA segment. We obtain the stationary distribution $\boldsymbol{\delta}_P = (0.620 \quad 0.358 \quad 0.022)$ with the seismic rates $\boldsymbol{\lambda}_P = (4.383 \quad 11.918 \quad 63.500)$. This means that the probability of mainshock earthquake frequency for a given year in the future (which is far enough from 2018) to be in a rare state (4.383) is about 62%, in a moderate state (11.918) is 34%, and in a frequent state (63.50) equals only 2%. Additionally, from the first row of TPM, we see that when the seismicity is in a low state, it then still remains in that state with a probability of 0.982, or it jumps to a medium state with a probability of 0, or it jumps to a high state with a probability of 0.018. This implies that a seismic level jump of two was feasible in this segment. The second and third rows of the TPM can be explained using the same technique. Moreover, let us consider the first column of Figure 5. From a time-dependent point of view, the low dynamic of the tectonic plate movements was recorded by the beginning of 1971 and, therefore, the seismic behavior jumped from low to high dynamics from 2006 to 2007. Thereafter, it remained in a moderate state from 2008 until 2018, as shown in the first row. Note that the period of 1971 to 2006 corresponds to the seismic quiescence: a lower rate period of mainshock earthquakes, as shown in the second row of Figure 5. The same analysis can be conducted for other segments and is omitted to save space. The property of seismic activity can jump two levels can also occur in the EO segment, shown by the representative result of the value $\gamma_{31} > 0$.

The Seismic Activity Analysis of Case 2

We now present the seismic activity analysis from a pair of two adjacent segments in the Sumatra megathrust zone (Case 2) wherein the possibility of spatial and/or time dependence may exist. For this purpose, we perform an analysis of four groups of adjacent segment pairs, namely, Aceh-Andaman with Nias-Simeuleu (AA-NS), Nias-Simeuleu with Mentawai-Siberut (NS-MS), Mentawai-Siberut with Mentawai-Pagai (MS-MP), and Mentawai-Pagai with Enggano (MP-EO). The pseudo-observation of data sets is calculated using Eq. 8 and is displayed graphically in the second row of Figure 6.

The Continuous Extension process using Eq. 9 is applied to obtain the continued discrete variables marked with an asterisk (*). For example, the continued variable for the AA segment is denoted as AA*, as well as for the others. The last row of Figure 6 shows the pseudo-observation scatter plots of the continued variable. Meanwhile, the graphical description of the Continuous Extension process, such as the Uniform distribution (0,1) uses, and the comparison of probability density (cumulative) functions between the discrete and its continued variables, is displayed in Figure 7. From now on, we focus on continued data, which will be used in estimating the parameters of the Copula-HMMs. Note that, before performing the estimation, three formal statistical tests, such as tests of spatial, stationarity, and serial time dependence, are first discussed.

The first statistical test procedure is adapted from Kruskal [55] in which the statistic tested is the significance of Kendall's τ. For the second test, the statistical test procedure is adapted from Bücher et al. [56] based on change-point detection in the Copulas, considering the following variants of the empirical process introduced by Csörgö and Horváth [57] and Gombay and Horváth [58]. For the third test, the statistical test procedure is worked out from two points of view, namely the bivariate and univariate series. For bivariate series, it is adapted from Genest and Remilland [59] based on the serial empirical Copula. Meanwhile, for univariate series, it is adapted from Ljung and Box [53] based on the sample autocorrelation function. The detailed formula and steps of three statistical tests are provided in the Appendix of this article.

In practice, the tests of three statistical are conducted using the copula R package provided by Hofert et al. [26]. These include cor(method=kendall), cpCopula(method=nonseg), and serial IndepTest(lag.max=finite number).

Results for the spatial dependence and stationarity tests are presented in Table 7, while the serial time dependence test is in Table 8. Note that, the *P*-values correspond to each statistical test, namely Z_τ (spatial dependence), S_T^C (stationarity), and $S_T^{\Pi^S}$(bivariate serial time dependence) (see the Appendix for details).

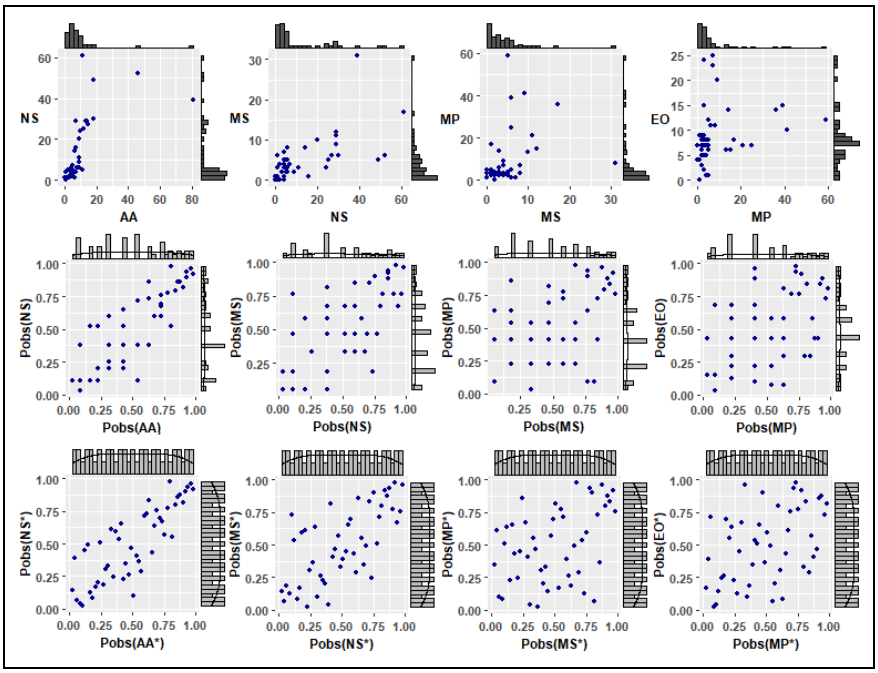

Figure 6: Rows correspond to (1) the bivariate frequency data, (2) the pseudo-observations of the bivariate discrete variable, and (3) the pseudo-observations of the bivariate continued variable. Columns correspond to the pairs of two adjacent segments.

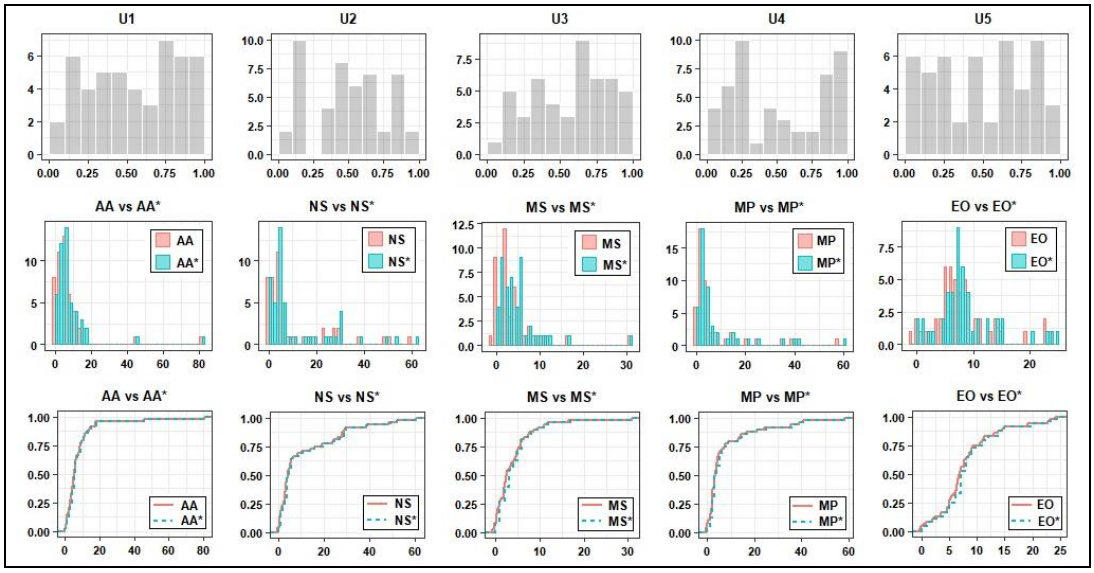

Figure 7: The first row corresponds to the Uniform distribution used on the discrete data of each segment. The second row illustrates the comparison of PDF between the discrete and its continued variable. The third row shows the comparison of CDF between the discrete and its continued variable. Meanwhile, columns correspond to the five segments studied.

Table 7: The test results of the dependence test using Z_τ and stationarity test using S_T^C.

Groups	Spatial Dependence Test				Stationarity Test		
	Kendall's τ	Z_τ	*P*-value	Decision of H_0	S_T^C	*P*-value	Decision of H_0
AA-NS	0.683	6.473	$9.6\text{x}10^{-11}$	Rejected	0.601	0.704	Accepted
NS-MS	0.524	4.909	$9.1\text{x}10^{-07}$	Rejected	0.391	0.778	Accepted
MS-MP	0.285	2.648	$8.1\text{x}10^{-03}$	Rejected	0.922	0.687	Accepted
MP-EO	0.323	3.039	$2.3\text{x}10^{-03}$	Rejected	0.411	0.709	Accepted

For the first test, the hypothesis is as follows: $H_0 : \tau = 0$ (independent) vs $H_1 : \tau \neq 0$ (dependent). As seen in Table 6 (part of the spatial dependence test), the small *P*-value provides evidence against independence for a pair of two data sets. Therefore, the spatial dependency of each pair analyzed exists.

For the second test, the null hypothesis, defined as a pair of two random variables, is stationary. It means that there exists a Copula C for the bivariate random variables. As shown in Table 7, part of the stationarity test and the large *P*-value (> 0.05) provide no evidence against the stationarity.

The small *P*-value for the third test, as displayed in Table 8, provides evidence against the null hypothesis, according to which the bivariate variable is serially time-independent. There is substantial evidence against serial time independence in the univariate component series, according to these results, which are consistent with those obtained from the perspective of the univariate test (Table 4).

Table 8: The test results of the bivariate serial time dependence test using Cramer-von Mises $S_T^{\Pi^s}$.

Segments	Cramer-von Mises Test $S_T^{\Pi^s}$		
	$S_T^{\Pi^s}$	*P*-value	Decision of H_0
AA	0.225	$4.9\text{x}10^{-04}$	Rejected
NS	0.294	$4.9\text{x}10^{-04}$	Rejected
MS	0.089	$1.4\text{x}10^{-03}$	Rejected
MP	0.147	$1.4\text{x}10^{-03}$	Rejected
EO	0.084	$4.9\text{x}10^{-04}$	Rejected

In summary, all the data sets of two adjacent segments show spatial and serial time dependence as well as stationarity features. These findings suggest that the spatial-time dependence modeling for the dataset can be constructed using Copula-HMMs approach.

To select the appropriate Copula-HMMs, the goodness of fit test is performed based on *P*-values criteria developed by Nasri et al. [30] that correspond with the Cramer-von Mises statistics $\mathcal{S}_T$ defined by Genest et al. [60], which is written as follows:

$$\mathcal{S}_T = T \int_{[0,1]^2} \big(D_T(\boldsymbol{u}) - \Pi(\boldsymbol{u})\big)^2 d\boldsymbol{u}. \tag{11}$$

The used variables in Eq. 11 are as follows.

Firstly, the empirical distribution $D_T(\boldsymbol{u})$ is formulated by $D_T(\boldsymbol{u}) = \frac{1}{T}\sum_{t=1}^{T} \prod_{j=1}^{2} \mathbf{1}\big(V_{T,j}(t) \leq u_j\big)$, where $V_{T,j}(t)$ is the pseudo-observations. Secondly, the independence Copula $\Pi(\boldsymbol{u}) = \prod_{j=1}^{2} u_j, \ \boldsymbol{u} \in [0,1]^2$, where u_j is the normalized ranks of the pseudo-observations. Furthermore, Thioub et al. [30] have built the package for this approach, namely the HMMcopula R-package. In this study, we applied that package with the number of bootstrap samples and the stopping criteria equal to 1000 and 0.0001, respectively.

The corresponding *P*-values are summarized in Table 9, together with the Cramer-von Mises statistics $\mathcal{S}_T$, and there are three comments from this table. Firstly, by looking at the biggest *P*-values (bold marks), the 2-state Gumbel-HMM is an appropriate model for the AA-NS group. Secondly, for the three groups remaining, Gaussian, Gumbel, and Frank Copula are the

selected models for NS-MS, MS-MP, and MP-EO groups, respectively. Thirdly, the two and three states of the Gaussian-HMMs (Gumbel-HMMs) have a significant *P*-value. However, based on the following review below, the estimated model parameters are independent of a number of states.

Let us consider the MS-MP group in Table 9. The statistical values of the three models of Gumbel-HMM with various states using the Cramer-von Mises $\mathcal{S}_T$ hardly differ from those of the *P*-values. Besides that, as can be seen in Table 10, in terms of Copula model parameter θ, namely Gumbel Copula (1.41507), 2-state Gumbel-HMM (1.41488; 1.41526), and 3-state Gumbel-HMM (1.415075; 1.415020; 1.415114), the parameter θ each state is likely to be worth ≈1.4150. Furthermore, the 90% confidence interval parameter of θ of the two models (i.e., 2-state Gumbel-HMM and 3-state Gumbel-HMM) tends to overlap, as can be seen in the 7th and 11th columns of Table 10. This condition suggests that the states of the two models are not well defined. Consequently, the 48 data pairs tend to be clustered only in one state. That is, for the 2-state Gumbel-HMM, all of the 48 data pairs are in the second state, meanwhile for the 3-state Gumbel-HMM, all of them are in the third state, as can be seen in "the number of components" row. According to the parsimony principle of mathematical modeling [19, 20], it is concluded that the Gumbel Copula is the final model for the MS-MP group.

Furthermore, the same explanation can be applied to the NS-MS group which is that for the final decision, the Gaussian Copula is still the best model.

The next results are related to the estimated parameters along with the 90% confidence intervals. Table 11 shows the estimated parameters of all groups based on the selected model. From the 7th through 11th rows (the parameter of θ and Kendall's τ), it suggests that the states of the AA-NS group are well defined. Additionally, from a spatial-time dependence modeling point of view, the seismic activity of 48 pairs of data earthquake frequency histories can be classified into two states based on the similarity of the appropriate Copula model. That is, states 1 and 2 correspond to the Gumbel Copula with parameters θ_1 = 2.037 and θ_2 = 3.629. Also, the numbers of components along with the Kendall's τ of states 1 and 2, which are 15 pairs of data (Kendall's τ = 0.509) and 33 pairs of data (Kendall's τ = 0.724) are obtained, respectively. These information are graphically represented in Figure 8 (bottom).

Table 9: Comparison of the fitted models on the basis of P-value. Also, the Cramer-von Mises $\mathcal{S}_T$ values are provided.

Copula-HMMs	m	Two Adjacent Segments Studied							
		AA-NS		NS-MS		MS-MP		MP-EO	
		$\mathcal{S}_T$	*P*-value	$\mathcal{S}_T$	*P*-value	$\mathcal{S}_T$	*P*-value	$\mathcal{S}_T$	*P*-value
Gaussian	1	0.068	0.053	0.032	**0.612**	0.031	0.540	0.014	0.998
	2	0.084	0.059	0.032	0.591	0.031	0.536	0.014	0.998
	3	0.083	< 0.05	0.032	0.592	0.031	0.562	0.014	0.998
Clayton	1	0.095	< 0.05	0.052	0.092	0.043	< 0.05	0.023	0.884
	2	0.124	0.075	0.048	0.314	0.068	0.558	0.023	0.875
	3	0.125	0.054	0.044	0.319	0.072	0.452	0.023	0.868
Gumbel	1	0.061	0.066	0.038	0.425	0.029	**0.588**	0.018	0.973
	2	0.069	**0.093**	0.038	0.384	0.029	0.565	0.018	0.976
	3	0.070	0.089	0.038	0.405	0.029	0.586	0.018	0.953
Frank	1	0.061	0.074	0.039	0.290	0.036	0.298	0.014	**1**
	2	0.060	0.061	n/a	n/a	n/a	n/a	n/a	n/a
	3	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Note: The notation of n/a denotes the value not available.

Table 10: Information regarding Copula-HMMs parameters for the NSMP and MS-MP group cases.

Parameters Model		NS-MS					MS-MP				
		2-state Gaussian-HMM		3-state Gaussian-HMM			2-state Gumbel-HMM		3-state Gumbel-HMM		
Value of cvm		0.032		0.032			0.02944		0.02944		
State		1	2	1	2	3	1	2	1	2	3
Parameter of $\boldsymbol{\theta}$	EM	0.7260	0.7280	0.7275	0.7273	0.7277	1.4149	1.4153	1.4151	1.4150	1.4151
	90% Conf. Limits	0.6553	0.6551	0.6763	0.6764	0.6763	1.2457	1.2454	1.2495	1.2491	1.2496
		0.8087	0.8084	0.8156	0.8156	0.8156	1.6666	1.6666	1.6375	1.6381	1.6381
Parameter of $\boldsymbol{\delta_C}$		0.4980	0.5020	0.3300	0.3060	0.3640	0.4880	0.5120	0.3300	0.3300	0.3400
Number of components		0	48	0	0	48	0	48	0	0	48
Estimated of Kendall's $\boldsymbol{\tau}$		0.518	0.519	0.5187	0.5185	0.5188	0.2932	0.2934	0.2933	0.2932	0.2933
Parameter of $\boldsymbol{\Gamma_C}$		$\begin{bmatrix} 0.498 & 0.502 \\ 0.498 & 0.502 \end{bmatrix}$		$\begin{bmatrix} 0.330 & 0.305 & 0.365 \\ 0.330 & 0.305 & 0.365 \\ 0.331 & 0.306 & 0.363 \end{bmatrix}$			$\begin{bmatrix} 0.487 & 0.513 \\ 0.490 & 0.510 \end{bmatrix}$		$\begin{bmatrix} 0.330 & 0.330 & 0.340 \\ 0.330 & 0.330 & 0.340 \\ 0.330 & 0.330 & 0.340 \end{bmatrix}$		

Table 11: The estimated parameter of the selected models for the four cases under study.

Selected and Parameters Model		Two Adjacent Segments Studied				
		AA-NS		NS-MP	MS-MP	MP-EO
		State 1	State 2			
Selected model		2-state Gumbel-HMM		Gaussian Copula	Gumbel Copula	Frank Copula
Value of cvm		0.069		0.032	0.029	0.014
Parameter of $\boldsymbol{\theta}$	EM	2.037	3.629	0.728	1.415	3.047
	90% Conf. Limits	1.931	2.024	0.716	1.335	2.510
		3.533	4.189	0.775	1.600	4.199
Estimated of Kendall's $\boldsymbol{\tau}$	EM	0.509	0.724	0.519	0.293	0.311
	90% Conf. Limits	0.499	0.550	0.507	0.255	0.268
		0.681	0.785	0.564	0.378	0.404
Number of components		15	33	48	48	48
Parameter of $\boldsymbol{\delta_C}$		0.415	0.585	1	1	1
Parameter of $\boldsymbol{\Gamma_C}$		$\begin{bmatrix} 0.036 & 0.964 \\ 0.691 & 0.309 \end{bmatrix}$		1	1	1

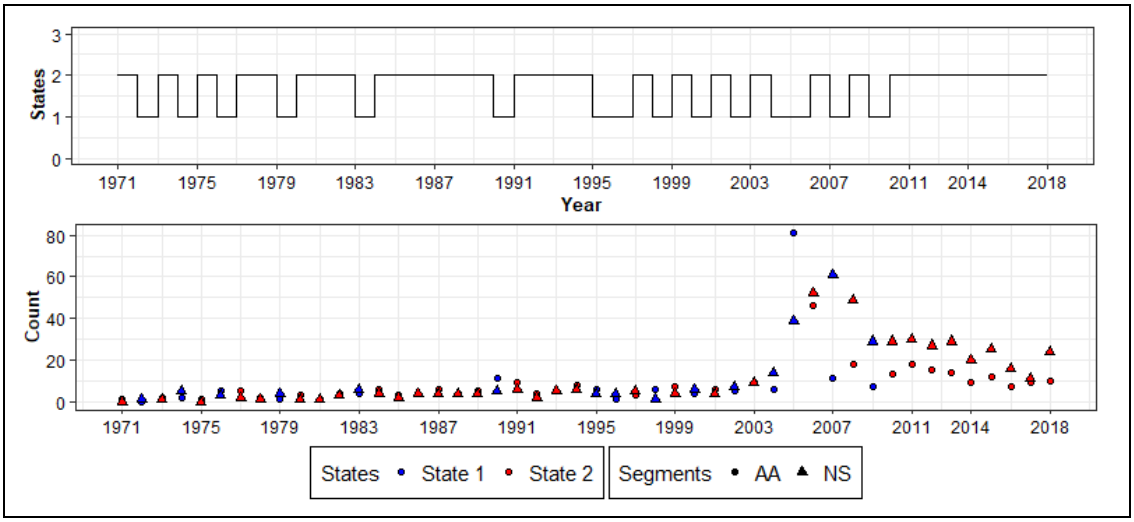

Figure 8: The AA-NS segments studied. The estimated states of seismic activity (top) and classification of the earthquakes are based on the state and its segment (bottom).

Subsequently, the estimated TPM $\boldsymbol{\Gamma_C}$ for the Gumbel-HMM with two states is determined in the last row of Table 10. From the first row of $\boldsymbol{\Gamma_C}$, when a data pair of AA-NS is in state 1, they remain in that state with a probability of only 0.036 or jump to state 2 with a probability of 0.964. Likewise, from the second row of $\boldsymbol{\Gamma_C}$, when a data pair of AA-NS is in state 2, they remain in that state with a probability of 0.309 or jump to state 1 with a probability equal to 0.691. Therefore, from a series point of view, the behavior of seismic activity to jump from one state to another has a greater probability than remaining in the same state, as graphically represented in Figure 8 (top). Meanwhile, from the stationary distribution, $\boldsymbol{\delta_C}$ = (0.415 0.585), the probability of seismic activity for a given year in the future (which is far enough from 2018) to be in state 1 is about 42%, while in state 2, it is equal to 58%.

CONCLUSIONS

In this paper, we have provided a seismic activity analysis in five segments of the Sumatra megathrust zone through an unobserved process study of tectonic plate movements. In practice, the analysis is worked out using two particular types of HMMs: Poisson-HMMs, which are used for each of the five segments independently analyzed (Case 1), and Copula-HMMs, which are used for a pair of two adjacent segments analyzed (Case 2). The analysis took into account the annual frequencies of mainshock earthquakes with $M_w \geq 4.6$ that occurred from 1971 to 2018.

The following are some conclusions from Case 1: each of the five independently analyzed segments. The seismic activity of each segment has serial time dependence, evidenced by the appropriate model for all five segments is the 3-state Poisson-HMM. The three states in this model represent the seismicity rates of mainshock earthquake frequency: rare, moderate, and frequent, which correspond to the level of the tectonic movement: low, medium, and high dynamics, respectively. Furthermore, based on the transition probability matrix, the seismic activity of AA and EO segments has the possibility to jump two levels, from low to high and from high to low, respectively. Therefore, these findings may lead geologists to further investigate how and why seismic activity in one partitioned megathrust zone can have different characteristics.

The following are some conclusions from Case 2, which is the seismic activity study of two adjacent segments (AA-NS, NS-MS, MS-MP, and MP-EO). The Continuous Extension

technique was first introduced to transform discrete variables into continuous ones due to the discrete property of seismicity data. From the four groups explored, it was found that for the AA-NS group, not only the serial time dependence that significantly exists in their earthquake histories but also spatial dependence, provided by the selected model that fits the data, namely the 2-state Gumbel-HMM. As for the remaining three groups, Nias-Simeulue with Mentawai-Siberut, Mentawai-Siberut with Mentawai-Pagai, and Mentawai-Pagai with Enggano, the appropriate models are Gaussian, Gumbel, and Frank Copulas, respectively. In this instance, the number of states represents the association of seismic activity between two adjacent segments, which is related to the association level between two adjacent tectonic plate dynamics.

From a practical point of view, the proposed model is rarely discussed in earthquake engineering studies. Therefore, these results will make a significant contribution to the seismic activity analysis in the earthquake engineering literature, especially for seismic issues in the Sumatra subduction zone.

ACKNOWLEDGMENTS

The authors sincerely wish to thank the journal editor and two reviewers who critically reviewed the manuscript and made valuable suggestions for its improvement. Furthermore, the authors acknowledge the funding provided by the P2MI research grant of Institut Teknologi Bandung, Indonesia 2021, with grant numbers: 62/IT1.C02/SK-TA/2021.

REFERENCES

1 Reid A (2015). "*History and Seismology in the Ring of Fire: Punctuating the Indonesian Past*". E-Book ISBN: 9789004288058, In Environment, Trade and Society in Southeast Asia Brill, 16pp. https://doi.org/10.1163/9789004288058_006

2 Irsyam M, Cummins PR, Asrurifak M, Faizal L, Natawidjaja DH. Widiyantoro S, Meilano I, Triyoso W, Rudiyanto A, Hidayati S, Ridwan M, Hanifa NR and Syahbana AJ (2020). "Development of the 2017 national seismic hazard maps of Indonesia". *Earthquake Spectra*, **36**: 112-136. https://doi.org/10.1177/8755293020951206

3 McCaffrey R (2009). "The tectonic framework of the Sumatran subduction zone". *Annual Review of Earth and Planetary Sciences*, **37**: 345-366. https://doi.org/10.1146/annurev.earth.031208.100212

4 Haridhi HA, Huang BS, Kuo-Liang W, Denzema D, Prasetyo RA and Chao-Shing L (2018). "A study of large earthquake sequences in the Sumatra subduction zone and its possible implications". *TAO: Terrestrial, Atmospheric and Oceanic Sciences*, **29**(6): 635-652. https://doi.org/10.3319/tao.2018.08.22.01

5 Meisl CS, Safaie S., Elwood KJ, Gupta R and Kowsari R (2006). "Housing reconstruction in northern Sumatra after the December 2004 great Sumatra earthquake and tsunami". *Earthquake Spectra*, **22**(S3): 777-802. https://doi.org/10.1193/1.2201668

6 Bothara J, Beetham D, Brunsdon D, Stannard M, Brown R, Hyland C, Lewis W, Miller S, Sanders R and Sulistio Y (2010). "General observations of effects of the 30th September 2009 Padang earthquake, Indonesia". *Bulletin of the New Zealand Society for Earthquake Engineering*, **43**(3): 143-173. https://doi.org/10.5459/bnzsee.43.3.143-173

7 Sieh K (2007). "The Sunda megathrust—past, present and future". *Journal of Earthquake and Tsunami*, **1**(01): 1-19. https://doi.org/10.1142/S179343110700002X

8 McCloskey J, Antonioli A, Piatanesi A, Sieh K, Steacy S, Nalbant S, Cocco M, Giunchi C, Huang J and Dunlop P (2008). "Tsunami threat in the Indian Ocean from a future megathrust earthquake west of Sumatra". *Earth and Planetary Science Letters*, **265**(1-2): 61–81. https://doi.org/10.1016/j.epsl.2007.09.034

9 Tabei T, Kimata F, Ito T, Gunawan E, Tsutsumi H, Ohta Y, Yamashina T, Soeda Y, Ismail N, Nurdin I, Sugiyanto D and Meilano I (2015). "Geodetic and geomorphic evaluations of earthquake generation potential of the northern Sumatran fault, Indonesia". *In International Symposium on Geodesy for Earthquake and Natural Hazards (GENAH)* **145**: 21-28. Springer, Cham. https://doi.org/10.1007/1345_2015_200

10 Natawidjaja DH (2018). "Updating active fault maps and sliprates along the Sumatran Fault Zone, Indonesia". *In IOP Conference Series: Earth and Environmental Science*, **118** (1): 012001. IOP Publishing. https://doi.org/10.1088/1755-1315/118/1/012001

11 Duputel Z, Kanamori H, Tsai VC, Rivera L, Meng L, Ampuero JP and Stock JM (2012). "The 2012 Sumatra great earthquake sequence". *Earth and Planetary Science Letters*, 351: 247-257. https://doi.org/10.1016/j.epsl.2012.07.017

12 Uphoff C, Rettenberger S, Bader M, Madden EH, Ulrich T, Wollherr S and Gabriel AA (2017). "Extreme scale multi-physics simulations of the tsunamigenic 2004 sumatra megathrust earthquake". *In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp.1-16. https://doi.org/10.1145/3126908.3126948

13 Panet I, Mikhailov V, Diament M, Pollitz F, King G, De Viron O, Holschneider M, Biancale R and Lemoine JM (2007). "Coseismic and post-seismic signatures of the Sumatra 2004 December and 2005 March earthquakes in GRACE satellite gravity". *Geophysical Journal International*, **171**(1): 177-190. https://doi.org/10.1111/j.1365-246x.2007.03525.x

14 Pesicek JD, Thurber CH, Zhang H, DeShon HR, Engdahl ER and Widiyantoro S (2010). "Teleseismic double-difference relocation of earthquakes along the Sumatra-Andaman subduction zone using a 3-D model". *Journal of Geophysical Research: Solid Earth*, **115**(B10): 1-20. https://doi.org/10.1029/2010jb007443

15 Orfanogiannaki K, Karlis D and Papadopoulos GA (2010). "Identifying seismicity levels via Poisson hidden Markov models". *Pure Applied Geophysics*, **167**(8–9): 919–931. https://doi.org/10.1007/s00024-010-0088-y

16 Orfanogiannaki K, Karlis D and Papadopoulos GA (2014). "Identification of temporal patterns in the seismicity of Sumatra using Poisson Hidden Markov models". *Research in Geophysics*, **4**(4969): 1-6. https://doi.org/10.4081/rg.2014.4969

17 Yip CF, Ng WL and Yau CY (2018). "A hidden Markov model for earthquake prediction". *Stochastic Environmental Research and Risk Assessment*, **32**(5): 1415-1434. https://doi.org/10.1007/s00477-017-1457-1

18 Zucchini W, MacDonald IL and Langrock R (2017). "*Hidden Markov Models for Time series: An Introduction Using R"*. eBook ISBN 9781315372488, Chapman and Hall/CRC. https://doi.org/10.1201/b20790-2

19 Joe H (1997). "*Multivariate Models and Multivariate Dependence Concepts*". eBook ISBN 9780367803896, CRC Press. https://doi.org/10.1201/9780367803896

20 Nelsen RB (2006). "*An Introduction to Copulas. 2nd Edition*". ISBN 978-0387-28659-4, New York: Springer. https://doi.org/10.1007/0-387-28678-0

21 Ogata Y (1999). "*Seismicity analysis through point-process modeling: A review*" page 471-507 in *Seismicity Patterns, their Statistical Significance and Physical Meaning*. Pageoph Topical Volumes, Birkhäuser, Basel. https://doi.org/10.1007/978-3-0348-8677-2_14

22 Can C, Ergun G and Gokceoglu C (2014). "Prediction of earthquake hazard by hidden Markov model (around Bilecik, NW Turkey)". *Open Geosciences*, **6**(3): 403-414. https://doi.org/10.2478/s13533-012-0180-1

23 Schweizer B and Sklar A (1983). "*Probabilistic Metric Spaces*". ISBN 0-444-00666-4, North-Holland Series in Probability and Applied Mathematics. https://www.academia.edu/26211504/B_Schweizer_A_Sklar_Probabilistic_metric_spaces_Elsevier_North_Holland_New_York_1983_pdf

24 Denuit M and Lambert P (2005). "Constraints on concordance measures in bivariate discrete data". *Journal of Multivariate Analysis*, **93**(1): 40-57. https://doi.org/10.1016/j.jmva.2004.01.004

25 Trivedi P and Zimmer D (2007). "Copula modeling: an introduction for practitioners". *Foundations and Trends in Econometrics*, **1**(1): 1-111. http://dx.doi.org/10.1561/0800000005

26 Hofert M, Kojadinovic I, Mächler M and Yan J (2019). "*Elements of copula modeling with R*". ISBN 978-3-319-89634-2. Springer. https://doi.org/10.1007/978-3-319-89635-9

27 Stevens WL (1950). "Fiducial Limits of the parameter of a discontinuous distribution". *Biometrika*, **37**: 117–129. https://doi.org/10.1093/biomet/37.1-2.117

28 Rizal J, Gunawan AY, Indratno SW and 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

29 Dempster AP, Laird NM and Rubin DB (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

30 Nasri BR, Rémillard BN and Thioub MY (2020). "Goodness-of-fit for regime-switching copula models with application to option pricing". *Canadian Journal of Statistics*, **48**(1): 79-96. https://doi.org/10.1002/cjs.11534

31 Thioub MY, Nasri BR, Pieugueu R and Rémillard BN (2020). "Package HMMcopula". https://cran.r-project.org/web/packages/HMMcopula/HMMcopula.pdf

32 Forney GD (1973). "The viterbi algorithm". *Proceedings of the IEEE*, **61**(3): 268-278. https://doi.org/10.1109/PROC.1973.9030

33 Zhang L and Singh VP (2019). "*Copulas and their applications in water resources engineering"* page 62-122, Online ISBN 9781108565103, Cambridge University Press. https://doi.org/10.1017/9781108565103.004

34 Patton AJ (2012). "A review of copula models for economic time series". *Journal of Multivariate Analysis*, **110**: 4-18. https://doi.org/10.1016/j.jmva.2012.02.021

35 Bárdossy A and Li J (2008). "Geostatistical interpolation using copulas". *Water Resources Research*, **44**(7): 1-15. https://doi.org/10.1029/2007wr006115

36 Kazianka, H., & Pilz, J. (2010). "Spatial interpolation using copula-based geostatistical models". *In GeoENV VII–Geostatistics for Environmental Applications*, Springer Dordrecht, **16**: 307-319. https://doi.org/10.1007/978-90-481-2322-3_27

37 Genest C, Gendron M and Bourdeau-Brien M (2009a). "The advent of copulas in finance". *The European Journal of Finance*, **15**(7-8): 609-618. https://doi.org/10.1080/13518470802604457

38 Aas K (2016). "Pair-copula constructions for financial applications: A review". *Econometrics*, **4**(4): 1-15. https://doi.org/10.3390/econometrics4040043

39 Sklar A (1959). "Functions de repartition an dimensions et leurs marges". *Publications de l'Institut de Statistique de l'Universite de Paris*, **8**: 229-231. https://ci.nii.ac.jp/naid/10011938360

40 Angus JE (1994). "The probability integral transform and related results". *SIAM Review*, **36**(4): 652-654. https://doi.org/10.1137/1036146

41 Joe H and Xu JJ (1996). "*The Estimation Method of Inference Functions for Margins for Multivariate Models*". Technical Report 166, Department of Statistics, University of British Columbia, 22pp. https://dx.doi.org/10.14288/1.0225985

42 Fermanian JD and Scaillet O (2005). "*Some statistical pitfalls in copula modelling for financical applications*", *E. Klein (Ed.), Capital Formation, Gouvernance and Banking*, 1-24. https://doi.org/10.2139/ssrn.558981

43 Genest C and Nešlehová J (2007). "A primer on copulas for count data". *ASTIN Bulletin: The Journal of the IAA*, **37**(2): 475-515. https://doi.org/10.2143/ast.37.2.2024077

44 Heinen A and Rengifo E (2007). "Multivariate autoregressive modeling of time series count data using copulas". *Journal of Empirical Finance*, **14**(4): 564-583. https://doi.org/10.1016/j.jempfin.2006.07.004

45 Nikoloulopoulos AK (2013). "On the estimation of normal copula discrete regression models using the continuous extension and simulated likelihood". *Journal of Statistical Planning and Inference*, **143**(11): 1923-1937. https://doi.org/10.1016/j.jspi.2013.06.015

46 Inouye DI, Yang E, Allen GI and Ravikumar P (2017). "A review of multivariate distributions for count data derived from the Poisson distribution". *Wiley Interdisciplinary Reviews: Computational Statistics*, **9**(3): e1398. https://doi.org/10.1002/wics.1398

47 van Stiphout T, Zhuang J and Marsan D (2012). "Seismicity declustering". *Community Online Resource for Statistical Seismicity Analysis*, **10**(1): 1-25. https://doi.org/10.5078/corssa-52382934

48 Mignan A and Woessner J (2012). "Estimating the magnitude of completeness for earthquake catalogs". *Community Online Resource for Statistical Seismicity Analysis*, 1-45. https://doi.org/10.5078/corssa-00180805

49 Wiemer S (2001). "A software package to analyze seismicity: ZMAP", *Seismological Research Letters*, **72**(3): 373-382. https://doi.org/10.1785/gssrl.72.3.373

50 Reasenberg P (1985). "Second-order moment of central California seismicity, 1969–1982". *Journal of Geophysical Research: Solid Earth*, **90**(B7): 5479-5495. https://doi.org/10.1029/jb090ib07p05479

51 Woessner J and Wiemer S (2005). "Assessing the quality of earthquake catalogues: Estimating the magnitude of completeness and its uncertainty". *Bulletin of the Seismological Society of America*, **95**(2): 684-698. https://doi.org/10.1785/0120040007

52 Schorlemmer D and Woessner J (2008). "Probability of detecting an earthquake". *Bulletin of the Seismological Society of America*, **98**(5): 2103-2117. https://doi.org/10.1785/0120070105

53 Ljung GM and Box GE (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

54 Schwarz G (1978). "Estimating the dimension of a model". *The Annals of Statistics*, **6**(2): 461-464. https://doi.org/10.1214/aos/1176344136

55 Kruskal WH (1958). "Ordinal measures of association". *Journal of the American Statistical Association*, **53**(284): 814-861. https://doi.org/10.1080/01621459.1958.10501481

56 Bücher A, Kojadinovic I, Rohmer T and Segers J (2014). "Detecting changes in cross-sectional dependence in multivariate time series". *Journal of Multivariate Analysis*, **132**: 111-128. https://doi.org/10.1016/j.jmva.2014.07.012

57 Csörgö M and Horváth L (1997). "*Limit Theorems in Change-Point Analysis*". ISBN 978-0-471-95522-1, Chichester: Wiley. https://www.wiley.com/en-us/Limit+Theorems+in+Change+Point+Analysis-p-9780471955221

58 Gombay E and Horváth L (1999). "Change-points and bootstrap". *Environmetrics: The official Journal of the International Environmetrics Society*, **10**(6): 725-736. https://doi.org/10.1002/(sici)1099-095x(199911/12)10:6%3C725::aid-env387%3E3.0.co;2-k

59 Genest C and Rémillard B (2004). "Test of independence and randomness based on the empirical copula process". *Test*, **13**(2): 335-369. https://doi.org/10.1007/bf02595777

60 Genest C, Rémillard B and Beaudoin D (2009b). "Goodness-of-fit tests for copulas: A review and a power study". *Insurance: Mathematics and Economics*, **44**(2): 199-213. https://doi.org/10.1016/j.insmatheco.2007.10.005

61 Kojadinovic I and Yan J (2011). "Tests of serial independence for continuous multivariate time series based on a Möbius decomposition of the independence empirical copula process". *Annals of the Institute of Statistical Mathematics*, **63**(2): 347-373. https://doi.org/10.1007/s10463-009-0257-x

APPENDIX

We list the detailed formula and steps of three statistical tests that are required in bivariate (time series) dependence modelling, namely spatial dependence, stationarity, and serial time dependence tests, respectively. Subsequently, in the rest of this appendix, we suppose that one has a random sample $\boldsymbol{x} = (x_1(t), x_2(t)) \in R^2$ which is the independent copies of bivariate random vector $\boldsymbol{X} = (X_1(t), X_2(t))$, $t \in \{1,2,\dots,T\}$.

Test of Serial Time Dependence

The following statistical procedures test of univariate serial dependence listed below can be followed:

1. Determine the hypothesis to be tested

   H_0: X_j are serially time-independent

   H_1: X_j are serially time-dependent $\quad j = 1,2.$

2. Determine an appropriate statistical test

$$Q_{LB}(h) = T(T+2)\sum_{t=1}^{h}\frac{\hat{\rho}^2(t)}{T-t}. \qquad \text{(A1)}$$

   The used variables in Eq. A1 are T is the sample size, h is the number of lags, and $\hat{\rho}^2(t)$ is the sample autocorrelation at lag t, which is calculated by

$$\hat{\rho}^2(t) = \frac{T\sum_{t=1}^{T-h}\left(x_j(t)-\bar{x_j}\right)\left(x_j(t+h)-\bar{x_j}\right)}{T-h\sum_{t=1}^{T-h}\left(x_j(t)-\bar{x_j}\right)^2}$$

   where $\bar{x_j} = \frac{1}{T}\sum_{t=1}^{T}x_j(t)$.

3. Calculate the approximate *P*-value for $Q_{LB}(h)$ test which is *P*-value $= \left(\Phi_{(\chi^2_{1-\alpha,h})}(Q_{LB}(h))\right)$. Here $\Phi_{(\chi^2_{1-\alpha,h})}$ denotes the CDF of the Chi-squared distribution with h degrees of freedom.

4. Decide whether the H_0 is accepted or rejected. The H_0 is rejected, if *P*-value for $Q_{LB}(h) < \alpha$.

The following statistical procedures test of bivariate serial time dependence listed below can be followed:

1. Determine the hypothesis to be tested

   H_0: $\boldsymbol{X}$ are serially time-independent

   H_1: $\boldsymbol{X}$ are serially time-dependent

2. Determine an appropriate statistical test

$$S_T^{\Pi^s} = \int_{[0,1]^p} T\left(C_{1:T}^s(\boldsymbol{u}) - \Pi(\boldsymbol{u})\right)^2 \mathrm{d}\boldsymbol{u}. \qquad \text{(A2)}$$

   The used variables in Eq. A2 are as follows:

$$C_{1:T}^s(\boldsymbol{u}) = \frac{1}{T}\sum_{i=1}^{T}\prod_{j=1}^{2}\mathbf{1}\left(G_{1:T}\left(X_{i+j-1}\right) \le u_j\right),$$

$$G_{1:T}(x) = \frac{1}{T+1}\sum_{j=1}^{T+1}\mathbf{1}\left(X_j \le x\right), \qquad x \in R.$$

$$\Pi(\boldsymbol{u}) = \prod_{j=1}^{2}u_j, \quad \boldsymbol{u} \in [0,1]^2.$$

3. Calculate the approximate *P*-value for $S_T^{\Pi^s}$. The detailed steps have been provided by Kojadinovic and Yan [61] using the permutation principle.

4. Decide whether the H_0 is accepted or rejected. The H_0 is rejected, if *P*-value for $S_T^{\Pi^s} < \alpha$.

Test of Spatial Dependence

In the setting under consideration, the procedure of spatial dependence test based on Kendall's τ listed below can be followed:

1. Determine the hypothesis to be tested

   H_0: $\boldsymbol{X}$ are spatially independent or $\tau(X_1, X_2) = 0$

   H_1: $\boldsymbol{X}$ are spatially dependent or $\tau(X_1, X_2) \neq 0$

2. Determine an appropriate statistical test

$$Z_{\tau(X_1,X_2)} = \frac{\tau(X_1,X_2)}{S_{\tau(X_1,X_2)}} = \frac{3(n_c - n_d)}{\sqrt{T(T-1)(2T+5)/2}} \qquad \text{(A3)}$$

   The used variables in Eq. A3 are as follows:

$$S_\tau = \sqrt{\frac{2(2T+5)}{9T(T-1)}}$$

$$= \sum_{i<j} \operatorname{sgn}\big(x_1(i) - x_1(j)\big) \operatorname{sgn}(x_2(i) - x_2(j))$$

$$\tau(X_1, X_2) = \frac{2}{T(T-1)}(n_c - n_d).$$

Variables n_c and n_d are the number of concordant and discordant pairs, respectively. That is, for $i < j$, if $\operatorname{sgn}(x_1(i) - x_1(j)) = \operatorname{sgn}(x_2(i) - x_2(j))$ then $(x_1(i), x_2(j))$ and $(x_1(j), x_2(j))$ is a concordance pair otherwise it is a discordance pair.

3. Calculate the approximate *P*-value for $Z_{\tau(X_1,X_2)}$, which is *P*-value $= 2\left(\Phi_{N(0,1)}\left(-\left|Z_{\tau(X_1,X_2)}\right|\right)\right)$. Here $\Phi_{N(0,1)}$ denotes the cumulative density function of Normal standart distribution.
4. Decide whether the H_0 is accepted or rejected. The H_0 is rejected, if *P*-value for $Z_\tau < \alpha$.

Test of Stationarity

The following statistical procedures test of stationarity listed below can be followed:

1. Determine the hypothesis to be tested

   H_0: $\boldsymbol{X}$ is a stretch from a stationarity time series

   H_1: $\boldsymbol{X}$ is a stretch from a non-stationarity time series
2. Determine an appropriate statistical test

$$S_T^C = \sup_{t\in[0,1]} \int_{[0,1]^2} \left(\mathbb{D}_T^C(t, \boldsymbol{u})\right)^2 \mathrm{d}C_{1:T}(\boldsymbol{u}) = \max_{1\le k\le n-1} \frac{1}{T}\sum_{i=1}^{T} \left(\mathbb{D}_T^C(k/T, \boldsymbol{U}_i^{1:T})\right)^2 \quad \text{(A4)}$$

   The used variables in Eq. A4 are as follows. Firstly, $\mathbb{D}_T^C(t, \boldsymbol{u}) = \sqrt{T}\lambda_T(0,t)\,\lambda_T(t,1)\left(C_{1:\lfloor Tt\rfloor}(\boldsymbol{u}) - C_{(\lfloor Tt\rfloor+1):T}(\boldsymbol{u})\right), (t, \boldsymbol{u}) \in [0,1]^3$, where, for any $0 \le a \le a' \le 1$, $\lambda_T(a, a') = (\lfloor Ta'\rfloor - \lfloor Ta\rfloor)/T$. Here, a notation $\lfloor . \rfloor$ denotes the floor function. The term of $\lambda_T(0,t)\,\lambda_T(t,1)$ is the cumulative sum change-point steps to which the class of tests belongs. Secondly, $C_{1:T}(\boldsymbol{u}) = \frac{1}{T}\sum_{i=1}^{T} \mathbf{1}(\boldsymbol{U}_i^{1:T} \le \boldsymbol{u})$, $\boldsymbol{u} \in [0,1]^2$ is the empirical Copula of $\boldsymbol{X}$. Thirdly, the sample of pseudo-observations $\boldsymbol{U}_i^{1:T}$ is defined by $\boldsymbol{U}_i^{1:T} = \frac{T}{T+1}\left(F_{1:T,1}(X_1(t)), F_{1:T,2}(X_2(t))\right)$ which is $F_{1:T,j}\left(X_j(t)\right)$ is the empirical CDF of $X_j(t)$.
3. Calculate the approximate *P*-value for S_T^C. The steps have been provided by Bücher et al. [57] using a resampling procedure.
4. Decide whether the H_0 is accepted or rejected. The H_0 is rejected, if *P*-value of $S_T^C < \alpha$.