IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 8, No. 5, October, 2021, pp. 793-804
Journal homepage: http://iieta.org/journals/mmep

The Application of Copula Continuous Extension Technique for Bivariate Discrete Data: A Case Study on Dependence Modeling of Seismicity Data

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

[1] Industrial and Financial Mathematics Research Group, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Bandung 40132, Indonesia
[2] Statistics Research Group, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Bandung 40132, Indonesia
[3] Geodesy Research Group, Faculty of Earth Science and Technology, Institut Teknologi Bandung, Bandung 40132, Indonesia

Corresponding Author Email: jose_rizal201@s.itb.ac.id

https://doi.org/10.18280/mmep.080516

Received: 8 November 2020
Accepted: 14 July 2021

Keywords:
continuous extension technique, dependence, copula model, Kendall's tau, random perturbation, earthquakes

ABSTRACT

The Copula approach for continuous variables is highly developed, while discrete ones are underdeveloped due to computational difficulties and sometimes algorithm failure to convergent. Therefore, providing an alternative method for discrete variables becomes an essential issue. In this paper, a simple method is proposed to answer the problem by applying the Continuous Extension Technique (CET). This is carried out by adding random independent perturbations in the form of either Uniform distribution $U(0,1)$ or $(U(0,1)-1)$, and the discrete variables are treated as continuous. Subsequently, a Copula model for resulted variables is estimated based on the Copula theory for continuous variables. This method is called a Copula continuous extension technique. Our analytic and simulation approaches show that both random perturbation forms produce the same Kendall's Tau measure and the selected Copula bivariate model. As illustrations, the proposed method is applied to the seismicity data obtained from the annual frequencies of earthquakes that occurred in the Sumatra megathrust of Indonesia, from January 1971 to December 2018, with magnitudes (M_w) of at least 4.6. Based on the selected Copula models, our simulations confirm the evidence of dependence seismic activity in each of the two adjacent large earthquake sources. These results provide new information regarding the seismicity behavior in the Sumatra megathrust.

1. INTRODUCTION

The Copula model approach becomes an alternative method in statistics, which is used in constructing the joint distribution of multivariate variables wherein dependence between marginals exists. Meanwhile, an application of the Copula model arises in a wide context, and several of the most commonly used are briefly mentioned, namely financial risk assessment by Zhang and Jiang [1], environmental sciences by Bhatti and Do [2], image processing by Dong et al. [3], health data by Ghahroodi et al. [4], and industrial problem by Wan and Li [5]. Furthermore, the Copula models used in the aforementioned cases are for continuous variables. Due to the non-uniqueness of the Copula model outside the range of marginal discrete distribution, the Copula for discrete variables is still not well developed [6]. In many real data applications, however, many phenomena are modeled based on the counting process, where the appropriate distributions are the discrete multivariate distributions. Therefore, these reasons motivate us to provide alternative procedures in Copula modeling for discrete variables.

One way to overcome this problem is by applying the Continuous Extension Technique (CET) to the marginal of the discrete variables. Moreover, the CET idea is to transform the marginal discrete variables X to become continuous, denoted by $X^\star$. This is worked out by summing the discrete random variable with a random perturbation taking values in $(0,1)$ (e.g., Uniform distribution, denotes $U(0,1)$) where the variables X and $U(0,1)$ are independent. Formally, it can be written as follows [7]:

$$X^\star = X + U(0,1). \quad (1)$$

According to Eq. (1), it is depicted that X is continued by $U(0,1)$. After applying the CET to the discrete variable (previous), a Copula model is estimated for continuous variables (new) by following the model procedure for continuous variables (referred to as Copula continuous extension technique).

The CET procedure has been studied further by Machado and Santos Silva [8] to extend quantile regression to count data and by Denuit and Lambert [9] to study concordance measures for dependent discrete data. The latter authors modified the CET, and Eq. (1) becomes $X^{\circledast} = X + (U(0,1) - 1)$. The detailed explanation is related to the idea of modifying the CET and its implications of the association measure (Kendall's Tau, denoted by τ) between the bivariate continuous (new) and the bivariate discrete (previous), that is, $\tau(X,Y) = \tau(X^{\circledast}, Y^{\circledast})$, can be seen in Denuit and Lambert [9].

Currently, the bivariate discrete variables (X,Y) are

considered, and are continued by (independent) considering two CET forms, $(X^{\star}, Y^{\star})$ according to Stevens [7] and $(X^{\circledast}, Y^{\circledast})$ according to Denuit and Lambert [9]. From each of the bivariate continuous, a Copula model can be constructed, as stated in Sklar's theorem [10]. From this situation, a question worthy to ask is whether the two Copula models produce the same model. When this happens, then one of two existing CET forms can be selected. Otherwise, further analysis regarding the selection of the appropriate CET model is needed. Since the result of this problem becomes the base for the next, this issue will be examined first. The examination is carried out using two approaches, namely analytic and simulation studies. In the simulation studies, five common bivariate discrete, namely Binomial, Geometric, Hypergeometric, Poisson, and Binomial Negative, are used to generate the dependence bivariate data. Meanwhile, for constructing the bivariate Copula, two families of Copula models, namely the Archimedean, i.e., Clayton, Gumbel, Frank, Joe, and Independent, and Elliptical, i.e., Gaussian and Student's, are applied.

In the present paper, the possible real case for the application was selected, which is the dependence modeling of seismicity data in the Sumatra megathrust subduction zone of Indonesia. The main reason for selecting this context is the characteristics of data to be analyzed appropriately with our research problems, which is a discrete multivariate variable. Moreover, to the best of our knowledge, the application of dependence modeling of seismicity data using the Copula model is still rare. Even though, this model is needed by seismologists to identify the dependencies of seismic activity in one restricted area [11].

The rest of the paper is organized as follows. The study materials and methods related to this work such as the theory of bivariate discrete distributions, association measure, and bivariate Copula theory are provided in Section 2. In Section 3, steps of simulation and corresponding analytical study are presented. Furthermore, the real data application on seismicity is described in Section 4. Finally, concluding results and recommendations for further study are discussed in Section 5.

2. MATERIALS AND METHODS

The use of a Copula model is relatively easy to implement in a bivariate case, but it becomes awkward when we implement in the multivariate case. Therefore, we restrict the discussion to the bivariate case, although generalization to higher dimensions is possible. In the next subsection, we address subjects such as bivariate discrete distribution, association measure, and bivariate Copula models. For the descriptions of bivariate discrete distribution, the books of Kocherlakota and Kocherlakota [12], Montgomery and Runger [13], were used as references. Meanwhile, for an association measure and bivariate Copula model, the books of Joe [14], Nelsen [15], and Hofert et al. [16] were used.

2.1 Bivariate discrete distributions

Here, we focus the discussion on five discrete bivariate distributions commonly used, namely Binomial, Geometric, Hypergeometric, Poisson, and Negative Binomial. Those distributions are derived from the Bernoulli distribution [17]. Furthermore, the relationship between the Bernoulli and the five aforementioned bivariate distributions is seen in Figure 1.

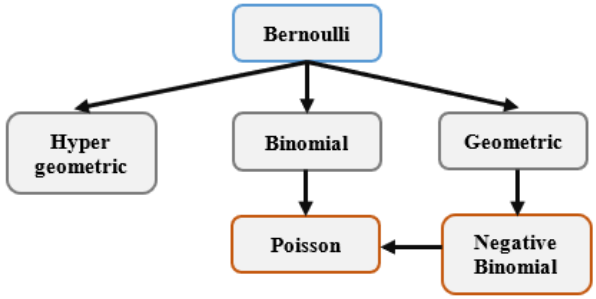

Figure 1. A family of bivariate distributions generated by the bivariate Bernoulli distribution (Adopted from Marshall and Olkin [17])

A common measure of the relationship between two random variables (X, Y) is the Pearson correlation coefficient (PCC) denoted by $\rho(X, Y)$. The value $\rho(X, Y)$ is the covariance of the two variables divided by the product of their standard deviations, which is formally stated as follows:

$$\rho(X, Y) = (\sigma_X \, \sigma_Y)^{-1} \, \mathrm{Cov}\,(X, Y). \tag{2}$$

As for the summary related to distribution notation-model parameters, joint probability mass function, and the Pearson correlation of the selected five bivariate discrete distributions are seen in Table 1.

2.2 Association measure

By far, the most familiar association measure between two random variables is calculated by the PCC defined in Eq. (2). However, it is noted that this approach has some limitations, namely only appropriate for detecting the linear dependencies between two variables and it is not invariant under strictly increasing non-linear transformation, which is, $\rho(X, Y) \neq \rho(F_X(x), G_Y(y))$. The variables $F_X(x)$ and $G_Y(y)$ are the cumulative distribution functions (CDFs) of the random variable of X and Y, respectively [18].

Those limitations motivate the development of a dependence measure for bivariate variables, i.e., Spearman's Rho and Kendall's Tau. Both measures are constructed based on the concept of concordance and discordance (Kruskal [19] and Lehmann [20]), which refers to the property that large values of one random variable tend to be associated with large values of the other random variable and small values of one random variable with small values of the other. Whereas discordance refers to large values of one random variable being associated with small values of the other.

To be more precise, suppose a pair of data is given by (x_i, y_i) and (x_j, y_j). We identify a concordance or discordance pair by: for $i \neq j$, if sign $(x_i - x_j)$ = sign $(y_i - y_j)$ then (x_i, y_i) and (x_j, y_j) is a concordance pair otherwise it is a discordance pair.

In the following discussions, we concentrate on Kendall's Tau measure which is denoted by "τ". By definition, Kendall's Tau measure is the probability of a concordance minus the probability of a discordance, which is formally written as follows:

$$\begin{aligned}\tau(X, Y) = {} & P[(X_1 - X_2)\,(Y_1 - Y_2) > 0] \\ & - P[(X_1 - X_2)\,(Y_1 - Y_2) < 0]\end{aligned} \tag{3}$$

Here, (X_1, Y_1) and (X_2, Y_2) are independent random vector and identically distributed from (X, Y). Also, it should be noted that Eq. (3) is appropriate for continuous random variables but not for discrete ones due to the presence of ties condition, that is, $P(\text{tie}) = P(X_1 = X_2 \text{ or } Y_1 = Y_2)$. By using that $P(\text{concordance}) + P(\text{discordance}) + P(\text{tie}) = 1$ then Kendall's Tau measure for the discrete bivariate variables (as commonly denotes τ_b) is written as follows [21]:

$$\tau_b(X, Y) = 4P[(X_1 - X_2)(Y_1 - Y_2) > 0] \\ -1 + P(X_1 = X_2 \text{ or } Y_1 = Y_2) \quad (4)$$

Table 1. Joint probability mass functions and the PCC formulation associated with the selected five bivariate discrete distributions adopted from Kocherlakota and Kocherlakota [12] and Montgomery and Runger [13]

Bivariate model	Distribution notation and model parameters	Joint probability mass function (PMF) formulation	$\boldsymbol{\rho}(X, Y)$
Binomial	BivBin $(n, k, l, p_{00}, p_{o1}, p_{10}, p_{11})$ $0 \le k, l \le n$ and $p_{00}, p_{o1}, p_{10}, p_{11} \ge 0$	$\sum_{\delta}^{\varepsilon} \frac{n! p_{00}^{n-(k+l)+\delta} p_{10}^{k-\delta} p_{01}^{l-\delta} p_{11}^{\delta}}{(n-(k+l)+\delta)!(k-\delta)!(l-\delta)!\delta!}$ $\varepsilon = \min(k, l)$ and $\delta = \max(k + l - n, 0)$	$\frac{n(p_{00}p_{11} - p_{10}p_{01})}{\sqrt{a}\sqrt{b}}$, $a = k(p_{10} + p_{11})(p_{00} + p_{01})$ $b = l(p_{01} + p_{11})(p_{00} + p_{10})$
Geometric	BivGeo(η_1, η_2, η_3) $0 < \eta_i < 1, i = 1,2$ $0 < \eta_3 \le 1$	$\eta_1^{x-1}\eta_2^{y-1}\eta_3^{z_1} - \eta_1^{x}\eta_2^{y-1}\eta_3^{z_2} - \eta_1^{x-1}\eta_2^{y}\eta_3^{z_3} + \eta_1^{x}\eta_2^{y}\eta_3^{z_4}$ $z_1 = \max(x-1, y-1)$, $z_2 = \max(x, y-1)$, $z_3 = \max(x-1, y)$, $z_4 = \max(x, y)$	$\frac{(1-\eta_3)\sqrt{\eta_1\eta_2}}{1-\eta_1\eta_2\eta_3}$
Hyper-geometric	BivHG$(N, (m_1, m_2), q)$ $q = x + y$, $N = m_1 + m_2$ and $q \le N$	$\frac{\binom{m_1}{x}\binom{m_2}{y}}{\binom{N}{q}}$	$-\sqrt{\frac{m_1}{N-m_1}\frac{m_2}{N-m_2}}$
Poisson	BivPoi$(\lambda_1, \lambda_2, \lambda_3)$ $\lambda_1, \lambda_2, \lambda_3 > 0$	$\exp\{-(\lambda_1 + \lambda_2 + \lambda_3)\}\frac{\lambda_1^x}{x!}\frac{\lambda_2^y}{y!}$ $\sum_{i=0}^{\min(x,y)}\binom{x}{i}\binom{y}{i} i! \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^i$	$\frac{\lambda_3}{\sqrt{\lambda_1+\lambda_3}\sqrt{\lambda_2+\lambda_3}}$
Binomial Negative	BivNB(κ, p_1, p_2) $\kappa > 0$ and $0 < p_i < 1$ such that $p_1 + p_2 < 1$	$\frac{\Gamma(x+y+\kappa)}{x!y!\Gamma(\kappa)} p_1^x p_2^y p_0^{\kappa}$ $p_0 = 1 - p_1 - p_2$	$\frac{p_1 p_2}{\sqrt{(p_0+p_1)(p_0+p_2)}}$

In addition, the association of two continuous random variables is measured using the bivariate Copula models approach, which is known as Kendall's Tau Copula τ_C. This is presented in the next subsection.

2.3 Bivariate Copula theory

To be self-contained, we provide a brief description of a bivariate Copula theory to sample-selection models. Literally, in statistics and probability theory, a bivariate Copula model is a two-dimensional joint CDFs based on a given marginal CDF. Therefore, the properties of the bivariate Copula models are analogous to properties of bivariate CDFs, that satisfy grounded and 2-increasing properties, see Joe [14], Nelsen [15], and Trivedi and Zimmer [18] for details.

The following theorem, known as Sklar's theorem [10], is the practical usefulness theorem of construction Copula models. According to that theorem, for random variables X and Y with respective marginal CDFs $F_X(x)$ and $F_Y(y)$, the bivariate distribution $H_{X,Y}(x, y)$ can be expressed as follows:

$$H_{X,Y}(x, y) = P(X \le x, Y \le y) \\ = C(F_X(x), F_Y(y): \theta), x, y \in R \quad (5)$$

where, θ is a parameter of the Copula model called the parameter of dependence, which describes the dependence between $F_X(x)$ and $F_Y(y)$. Estimation θ by the value $\hat{\theta}$ can be determined by maximizing the log-likelihood function of the pdf Copula model as follows [22]:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{T} \log c(F_X(x_i), F_Y(y_i); \theta) \quad (6)$$

Based on the Integral Transform probability theorem (Angus [23]), the marginal distribution of variables (X, Y) follows the Uniform distribution ranging from 0 to 1. Therefore, by denotes $F_X(x)$ and $F_Y(y)$ with u_1 and u_2, respectively, we can rewrite the Eq. (5) as follows:

$$C(F_X, F_Y: \theta) = C(u_1, u_2: \theta), u_1, u_2 \in [0,1] \quad (7)$$

There are a number of Copula models that have been widely explored in many articles and books. In this paper, models are chosen to be the Archimedean Copula family (i.e., Clayton, Gumbel, Frank, Joe, and Independent) and the Elliptical Copula family (i.e., Gaussian and Student's) which were often used in practice. Here, we write those Copula models in terms of random variables U_1 and U_2 that have standard to denotes the Uniform marginal distribution, as can be seen in Table 2.

3. ANALYTICAL AND SIMULATION STUDIES

Let (X, Y) be bivariate discrete random variables. We reconsider two random independent perturbation forms of the CET process for marginal bivariate discrete variables, i.e., $X^{\star} = X + U(0,1)$ [7] and $X^{\circledast} = X + (U(0,1) - 1)$ [9]. There is analog for the discrete variable Y, but now replacing $U(0,1)$ by $V(0,1)$. Here, the variable $U(0,1)$ and $V(0,1)$ are independent. Subsequently, we analytically prove that for a given bivariate discrete variables (X, Y) the two CET results, i.e., $(X^{\star}, Y^{\star})$ and $(X^{\circledast}, Y^{\circledast})$, produce the same Copula model and its parameter. Therefore, some simulations are carried to confirm our analytical approach. The process includes the following.

Suppose $(F_{X^{\star}}, F_{Y^{\star}})$ and $(F_{X^{\circledast}}, F_{Y^{\circledast}})$ are the CDFs for $(X^{\star}, Y^{\star})$ and $(X^{\circledast}, Y^{\circledast})$, respectively. We will first examine that if two bivariate continuous, $(X^{\star}, Y^{\star})$ and $(X^{\circledast}, Y^{\circledast})$ have the same behavior of the CDFs, i.e., $(F_{X^{\star}}, F_{Y^{\star}}) = (F_{X^{\circledast}}, F_{Y^{\circledast}})$,

and Kendall's Tau measure, i.e., $\tau(\mathrm{X}^{\star},\mathrm{Y}^{\star})=\tau(\mathrm{X}^{\circledast},\mathrm{Y}^{\circledast})$, then they will have not only the same Copula model but also its parameter due to the presence of the unique property of the Copula model for continuous variables [10].

Table 2. Formulation of some single parameter Copula model of the selected Archimedean Copulas family with their generators function and the selected Elliptical Copulas family, i.e., Gaussian Copula and Students's Copula

Copula model	Function $C(u_1,u_2;\theta)$	Generator $\phi(t;\theta)$	Parameter range(θ)	Kendall's tau of the copula C
Independence	$u_1\,u_2$	$-\log(t)$	n/a	0
Clayton	$\left(u_1^{-\theta}+u_2^{-\theta}-1\right)^{-1/\theta}$	$\theta^{-1}\left(t^{-\theta}-1\right)$	$(0,\infty)$	$\frac{\theta}{\theta+2}$
Gumbel Hougaard	$\exp\left(-\left(\tilde{u}_1^{\theta}+\tilde{u}_2^{\theta}\right)^{1/\theta}\right)$; $\tilde{u}_j=-\log u_j$	$(-\log t)^{\theta}$	$[1,\infty)$	$\frac{\theta-1}{\theta}$
Frank	$-\frac{1}{\theta}\log\left(1+\frac{u_1{}^{*}\,u_2{}^{*}}{\exp(-\theta)-1}\right)$; $u_j{}^{*}=e^{(-\theta u_j-1)}$	$-\log\left(\frac{\exp(-\theta t)-1}{\exp(-\theta)-1}\right)$	$(-\infty,\infty)$	$1-\frac{4}{\theta}[1-D_1(\theta)]$
Joe	$1-[u_1^{*}+u_2^{*}-u_1^{*}\,u_2^{*}]^{1/\theta}$; $u_j{}^{*}=\left(1-u_j\right)^{\theta}$	$-\log\left(1-(1-t)^{\theta}\right)$	$[1,\infty)$	$1-\sum_{k=1}^{\infty}\frac{4}{h(\theta,k)}$
Gaussian	$\Phi_G[\Phi^{-1}(u_1),\Phi^{-1}(u_2);\theta]$	n/a	$(-1,1)$	$\frac{2}{\pi}\arcsin(\theta)$
Student's	$t_{2,\nu}[t_{\nu}^{-1}(u_1),t_{\nu}^{-1}(u_2);\theta]$; $\nu\in(2,\infty)$	n/a	$[-1,1]$	$\frac{2}{\pi}\arcsin(\theta)$

Notes: 1. The quantity D_1 is the Debye function of order one, defined by $D_1(x)=1/x\int_0^x t/(e^t-1)\,dt$, $t\in(0,\infty)$, 2. $h(\theta,k)=\left(k(\theta k+2)(\theta(k-1)+2)\right)$, and 3. n/a denotes that the mentioned item is not available.

Before proceeding to show $(F_{\mathrm{X}^{\star}},F_{\mathrm{Y}^{\star}})=\left(F_{\mathrm{X}^{\circledast}},F_{\mathrm{Y}^{\circledast}}\right)$, a relation of $F_{\mathrm{X}^{\star}}$ with $F_{\mathrm{X}^{\circledast}}$ and $F_{\mathrm{Y}^{\star}}$ with $F_{\mathrm{Y}^{\circledast}}$ needs to be determined. In addition, for all (x,y) elements of the domain $(\mathrm{X}^{\circledast},\mathrm{Y}^{\circledast})$, there is $(x+1,\mathrm{y}+1)$ elements of the domain $(\mathrm{X}^{\star},\mathrm{Y}^{\star})$ such that $F_{\mathrm{X}^{\circledast}}(x)=P(\mathrm{X}^{\circledast}\le x)=P(\mathrm{X}_i^{\star}-1\le x)=P(\mathrm{X}_i^{\star}\le x+1)=F_{\mathrm{X}^{\star}}(x+1)$. The same statement also holds for $F_{\mathrm{Y}^{\circledast}}(x)=F_{\mathrm{Y}^{\star}}(y+1)$.

As already stated, we will prove $(F_{\mathrm{X}^{\star}},F_{\mathrm{Y}^{\star}})=\left(F_{\mathrm{X}^{\circledast}},F_{\mathrm{Y}^{\circledast}}\right)$ by showing that $(F_{\mathrm{X}^{\star}},F_{\mathrm{Y}^{\star}})\subseteq\left(F_{\mathrm{X}^{\circledast}},F_{\mathrm{Y}^{\circledast}}\right)$ and $(F_{\mathrm{X}^{\star}},F_{\mathrm{Y}^{\star}})\supseteq\left(F_{\mathrm{X}^{\circledast}},F_{\mathrm{Y}^{\circledast}}\right)$.

We start the investigation for $(F_{\mathrm{X}^{\star}},F_{\mathrm{Y}^{\star}})\subseteq\left(F_{\mathrm{X}^{\circledast}},F_{\mathrm{Y}^{\circledast}}\right)$. Suppose that $(a_0,b_0)\in(F_{\mathrm{X}^{\star}},F_{\mathrm{Y}^{\star}})$, by using the CDF definition, there is (x_0,y_0), which belongs to the domain of $(\mathrm{X}^{\star},\mathrm{Y}^{\star})$ such that $a_0=F_{\mathrm{X}^{\star}}(x_0)=F_{\mathrm{X}^{\circledast}}(x_0-1)$ and $b_0=F_{\mathrm{Y}^{\star}}(y_0)=F_{\mathrm{Y}^{\circledast}}(y_0-1)$. Thus, $a_0=F_{\mathrm{X}^{\circledast}}(x_0-1)$ and $b_0=F_{\mathrm{Y}^{\circledast}}(y_0-1)$ where (x_0-1,y_0-1) belongs to the domain of $(\mathrm{X}^{\circledast},\mathrm{Y}^{\circledast})$ and thus $(a_0,b_0)\in\left(F_{\mathrm{X}^{\circledast}}(\mathrm{X}^{\circledast}),F_{\mathrm{Y}^{\circledast}}(\mathrm{Y}^{\circledast})\right)$. As a conclusion $(F_{\mathrm{X}^{\star}},F_{\mathrm{Y}^{\star}})\subseteq\left(F_{\mathrm{X}^{\circledast}},F_{\mathrm{Y}^{\circledast}}\right)$.

Next, we check for $(F_{\mathrm{X}^{\star}},F_{\mathrm{Y}^{\star}})\supseteq\left(F_{\mathrm{X}^{\circledast}},F_{\mathrm{Y}^{\circledast}}\right)$. We work out in similar way. Let (a_1,b_1) be the element of $\left(F_{\mathrm{X}^{\circledast}},F_{\mathrm{Y}^{\circledast}}\right)$. This means, there is (x_1,y_1), which belongs to domain of $(\mathrm{X}^{\circledast},\mathrm{Y}^{\circledast})$ such that $a_1=F_{\mathrm{X}^{\circledast}}(x_1)=F_{\mathrm{X}^{\star}}(x_1+1)$ and $b_1=F_{\mathrm{Y}^{\circledast}}(y_1)=F_{\mathrm{Y}^{\star}}(y_1+1)$. In other words, $a_1=F_{\mathrm{X}^{\star}}(x_1+1)$ and $b_1=F_{\mathrm{Y}^{\star}}(y_1+1)$ where (x_1+1,y_1+1) belongs to domain of $(\mathrm{X}^{\star},\mathrm{Y}^{\star})$. So, we have that $(a_1,b_1)\in\left(F_{\mathrm{X}^{\star}}(\mathrm{X}^{\star}),F_{\mathrm{Y}^{\star}}(\mathrm{Y}^{\star})\right)$ and thus, $(F_{\mathrm{X}^{\star}},F_{\mathrm{Y}^{\star}})\supseteq\left(F_{\mathrm{X}^{\circledast}},F_{\mathrm{Y}^{\circledast}}\right)$.

By combining the two results above, i.e., $(F_{\mathrm{X}^{\star}},F_{\mathrm{Y}^{\star}})\subseteq\left(F_{\mathrm{X}^{\circledast}},F_{\mathrm{Y}^{\circledast}}\right)$ and $(F_{\mathrm{X}^{\star}},F_{\mathrm{Y}^{\star}})\supseteq\left(F_{\mathrm{X}^{\circledast}},F_{\mathrm{Y}^{\circledast}}\right)$, we therefore have $(F_{\mathrm{X}^{\star}},F_{\mathrm{Y}^{\star}})=\left(F_{\mathrm{X}^{\circledast}},F_{\mathrm{Y}^{\circledast}}\right)$.

The second examination is related to Kendall's Tau measure of $(\mathrm{X}^{\star},\mathrm{Y}^{\star})$ and $(\mathrm{X}^{\circledast},\mathrm{Y}^{\circledast})$. Let $(\mathrm{X}_1,\mathrm{Y}_1)$ and $(\mathrm{X}_2,\mathrm{Y}_2)$ be independent copies of bivariate discrete (X,Y). We assume that for $i=1,2$ holds: (i) X_i and Y_i are continued by the method of Stevens [7] or Denuit and Lambert [9] (ii) the Uniform distribution U_i and V_i are independent. Under these conditions and according to Eq. (3), Kendall's Tau measure of $(\mathrm{X}^{\circledast},\mathrm{Y}^{\circledast})$ is written as follows: $\tau(\mathrm{X}^{\circledast},\mathrm{Y}^{\circledast})=P\left[\left(X_1^{\circledast}-X_2^{\circledast}\right)\left(Y_1^{\circledast}-Y_2^{\circledast}\right)>0\right]-P\left[\left(X_1^{\circledast}-X_2^{\circledast}\right)\left(Y_1^{\circledast}-Y_2^{\circledast}\right)<0\right]$.

Note that, we write the expression of $\mathrm{X}_i^{\circledast}=\mathrm{X}_i+(\mathrm{U}_i(0,1)-1)$ to become $\mathrm{X}_i^{\circledast}=\mathrm{X}_i^{\star}-1$, similarly for the variable $\mathrm{Y}_i^{\circledast}$. Thus, we get

$$\begin{aligned}\tau(\mathrm{X}^{\circledast},\mathrm{Y}^{\circledast})&=P\left[\left((X_1^{\star}-1)-(X_2^{\star}-1)\right)\left((Y_1^{\star}-1)-(Y_2^{\star}-1)\right)>0\right]-P\left[\left((X_1^{\star}-1)-(X_2^{\star}-1)\right)\left((Y_1^{\star}-1)-(Y_2^{\star}-1)\right)<0\right]\\&=P[(X_1^{\star}-X_2^{\star})\,(Y_1^{\star}-Y_2^{\star})>0]-P[(X_1^{\star}-X_2^{\star})\,(Y_1^{\star}-Y_2^{\star})<0]=\tau(\mathrm{X}^{\star},\mathrm{Y}^{\star})\end{aligned}\tag{8}$$

The last expression of Eq. (8) confirmed that $\tau(\mathrm{X}^{\star},\mathrm{Y}^{\star})=\tau(\mathrm{X}^{\circledast},\mathrm{Y}^{\circledast})$. In addition, Denuit and Lambert [9] have stated that for a given bivariate discrete (X,Y), then $\tau(\mathrm{X},\mathrm{Y})=\tau(\mathrm{X}^{\circledast},\mathrm{Y}^{\circledast})$. Hence, based on those results, we conclude that $\tau(\mathrm{X},\mathrm{Y})=\tau(\mathrm{X}^{\star},\mathrm{Y}^{\star})=\tau(\mathrm{X}^{\circledast},\mathrm{Y}^{\circledast})$. This means the two random independent perturbation forms of the CET process in bivariate discrete did not change Kendall's Tau measure.

To summarize the analytic studies, in setting consideration for our research problem, two necessary conditions are provided, namely $(F_{\mathrm{X}^{\star}},F_{\mathrm{Y}^{\star}})=\left(F_{\mathrm{X}^{\circledast}},F_{\mathrm{Y}^{\circledast}}\right)$ and $\tau(\mathrm{X}^{\star},\mathrm{Y}^{\star})=\tau(\mathrm{X}^{\circledast},\mathrm{Y}^{\circledast})$, such that $C(F_{\mathrm{X}^{\star}},F_{\mathrm{Y}^{\star}}:\theta^{\star})=C\left(F_{\mathrm{X}^{\circledast}},F_{\mathrm{Y}^{\circledast}}:\theta^{\circledast}\right)$ with $\theta^{\star}=\theta^{\circledast}$. In other words, they have the same Copula model and its parameter.

Subsequently, some simulations are presented to confirm the analytical approach. In the simulations, data are drawn from five bivariate discrete (see Table 1). Furthermore, a sample of size N = 200 is generated by choosing a set of parameters corresponding to the characteristics of each distribution model. The characteristics are referred to as low PCC (below of 0.5) and high (above of 0.5) except for Hypergeometric. Therefore, there are nine possible cases to be studied. The procedure simulation listed below can be followed:

1. Generate a vector (X,Y) of bivariate data from one of the five bivariate discrete distributions.

2. Apply Continuous Extension Technique (CET) to marginals bivariate discrete distributions X and Y with three different random perturbation forms, i.e., $U(0,1)$ (Control), and $(U(0,1)-1)$ (Treatment 1): (Stevens, [7]), (Denuit and Lambert [9]), respectively, plus $(U(0,1)+1)$ (Treatment 2) so that yielding three bivariate continuous (new), namely $(X^{\star}, Y^{\star})$, $(X^{\circledast}, Y^{\circledast})$, and $(X^{\star}, Y^{\star})$.
3. Identify the marginal density function of the three bivariate continuous variables.
4. Estimate Kendall's Tau Empiric (τ_E) of $(X^{\star}, Y^{\star})$, $(X^{\circledast}, Y^{\circledast})$, and $(X^{\star}, Y^{\star})$ using Eq. (3).
5. Construct the bivariate Copula model of the $(X^{\star}, Y^{\star})$, $(X^{\circledast}, Y^{\circledast})$, and $(X^{\star}, Y^{\star})$. Here, the seven types of bivariate Copula models as display in Table 2 are evaluated altogether.
6. Estimate Kendall's Tau Copula (τ_C) of $(X^{\star}, Y^{\star})$, $(X^{\circledast}, Y^{\circledast})$, and $(X^{\star}, Y^{\star})$.
7. Repeat the following steps 100 times and report the τ_E, selected Copula model, and τ_C.

It must be noted that for the construction continuous Copula model, a two-step maximum likelihood (TSML) procedure is commonly used. That is, the marginals are estimated in the first step, and then as the second step, the dependence parameter is estimated using the selected family of Copula models based on the CDF of the inferred marginal distributions. This procedure is known as the Inference Function for Marginals (IFM) [24].

In the third simulation step, marginal probability distributions of continuous variable are identified by fitting with eight distribution models, namely Normal, Logistic, Cauchy, Exponential, LogNormal, Gamma, Weibull, and Gumbel.

The notation, formulation, and parameters of each continuous probability function are presented briefly in Table 3. The standard approach to estimate the parameters models ($\boldsymbol{\omega}$) of each distribution is the maximum likelihood method, which requires the maximization of the log-likelihood function that represented as follows [25]:

$$\mathcal{L}(\boldsymbol{\omega}|x_1, \dots, x_T) = \sum_{i=1}^{T} \log f\big((x_i); \boldsymbol{\omega}\big) \tag{9}$$

Table 3. Probability and cumulative distribution functions associated with the selected eight continuous distribution models

Univariat model	Probability density function $f(x; \boldsymbol{parameters})$	Cumulatif density function $F(x; \textbf{parameters})$	Domain and parameters range
$\textbf{Normal}(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\,\sigma}\right)\right)$	$x, \mu \in (-\infty, +\infty)$ $\sigma^2 \in (0, +\infty)$
$\textbf{Logistic}(\mu, s)$	$\frac{\exp\left(-\frac{(x-\mu)}{s}\right)}{s\left(1 + \exp\left(-\frac{(x-\mu)}{s}\right)\right)^2}$	$\frac{1}{1 + \exp\left(-\frac{(x-\mu)}{s}\right)}$	$x, \mu \in (-\infty, +\infty)$ $s \in (0, +\infty)$
$\textbf{Cauchy}(x_0, \gamma)$	$\frac{1}{\pi\gamma}\left(\frac{\gamma^2}{(x-x_0)^2 + \gamma^2}\right)$	$\frac{1}{\pi}\arctan\left(\frac{x-x_0}{\gamma}\right) + \frac{1}{2}$	$x, x_0 \in (-\infty, +\infty)$ $\gamma \in (0, +\infty)$
$\textbf{Exponential}(\xi)$	$\xi \exp(-\xi x)$	$1 - \exp(-\xi x)$	$x \in [0, +\infty)$ $\xi \in (0, +\infty)$
$\textbf{Lognormal}(\mu, \sigma^2)$	$\frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$	$\frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{\ln x - \mu}{\sqrt{2}\,\sigma}\right)\right)$	$x, \sigma^2 \in (0, +\infty)$ $\mu \in (-\infty, +\infty)$
$\textbf{Gamma}(h, \vartheta)$	$(\Gamma(h)\,\vartheta^h)^{-1}\, x^{h-1} e^{-\vartheta/x}$	$\frac{1}{\Gamma(h)}\,\gamma\left(h, \frac{x}{\vartheta}\right)$	$x, h, \vartheta \in (0, +\infty)$
$\textbf{Weibull}(\varrho, k)$	$\frac{k}{\varrho}\left(\frac{x}{\varrho}\right)^{k-1} \exp\left(-\left(\frac{x}{\varrho}\right)^k\right)$	$1 - \exp\left(-\left(\frac{x}{\varrho}\right)^k\right)$	$x \in [0, +\infty)$ $k, \varrho \in (0, +\infty)$
$\textbf{Gumbel}(\mu, \beta)$	$\frac{1}{\beta} \exp\left(-\left(\frac{x-\mu}{\beta} + e^{-\frac{x-\mu}{\beta}}\right)\right)$	$\exp\left(-e^{-\frac{x-\mu}{\beta}}\right)$	$x, \mu \in (-\infty, +\infty)$ $\beta \in (0, +\infty)$

Note: $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{t^2} dt$.

The goodness-of-fit quantification level between the given data set (as information) and a theoretical model fitted is required in steps 3 and 4. Due to the theoretical model parameters (Tables 2 and 3) are estimated using the maximum likelihood, then the information-theoretical, such as the Akaike and Bayesian Information Criteria (AIC and BIC , respectively) are used for model selection, as recommended by Akaike [26] and Schwarz [27]. These criteria are defined as follows:

$$\mathrm{AIC}(l) = -2\mathcal{L}_l + 2p(l). \tag{10}$$

$$\mathrm{BIC}(l) = -2\mathcal{L}_l + p(l)\log \mathrm{T}. \tag{11}$$

Variable $\mathcal{L}_l$ is the value for the maximum likelihood of the fitted model, $\mathrm{p}(l)$ is the number of parameters, and variable T is the number of observations. Furthermore, the smaller values of $\mathrm{AIC}(l)$ and $\mathrm{BIC}(l)$ means that the selected model shows better agreement with the data set.

In the present paper, the simulation and visualizations are carried out using the R program. Furthermore, several packages in the R program that correspond to this study include "extraDistr" (Wolodzko [28]), "BivGeo" (de Oliveira and Achcar [29]), "bivariate" (Spurdle [30]), "fitdistrplus" (Delignette-Muller and Dutang [31]), "RMKdiscrete" (Kirkpatrick [32]), "copula" (Hofert et al. [33]), and "VineCopula" (Schepsmeier et al. [34]). As information, steps to construct the bivariate Copula model for continuous variables have been provided by Xu et al. [35]. Meanwhile, Hofert et al. [16] have provided the elements of Copula Modeling with R program.

To summarize the simulation studies, there are two main results from 100 iterations performed. Firstly, the comparison of the frequency histograms of dependence parameter, the

Kendall's Tau Empiric (τ_E) of three continuous bivariates, that is, $(X^{\star}, Y^{\star})$, $(X^{\circledast}, Y^{\circledast})$, and $(X^{\star\star}, Y^{\star\star})$, with the random perturbation forms $U(0,1)$ as control, $(U(0,1)-1)$ as treatment 1, and $(U(0,1)+1)$ as treatment 2, respectively). These are shown as the same histogram (the left-hand panels of Figure 2). In other words, the three data sets of 100 iterations performed are the same.

Secondly, as seen in the right-hand panels of Figure 2, a comparison of the frequency histograms of dependence parameter, which is, Kendall's Tau Copula τ_C of $(X^{\star}, Y^{\star})$, $(X^{\circledast}, Y^{\circledast})$, and $(X^{\star\star}, Y^{\star\star})$ are shown as the same histogram. In other words, each of 100 iterations performed the same as the selected Copula model.

The simulation results are in line with the findings of the analytic studies. Accordingly, the CET form of Denuit and Lambert [9] and Stevens [7] gave the same conclusion, i.e., produced the same Copula model and its parameter. Therefore, in practice, one of two existing random independent perturbation forms of the CET process can be selected in modifying the Copula methods for bivariate discrete variables.

The next discussion is related to the illustration of our proposed technique on the dependence modeling of seismicity data. Hence, to assist in explaining the illustration for the seismologists, we will elaborate on the CET process of bivariate discrete (previous) and the steps of the Copula modeling for the bivariate continuous (new) in detail.

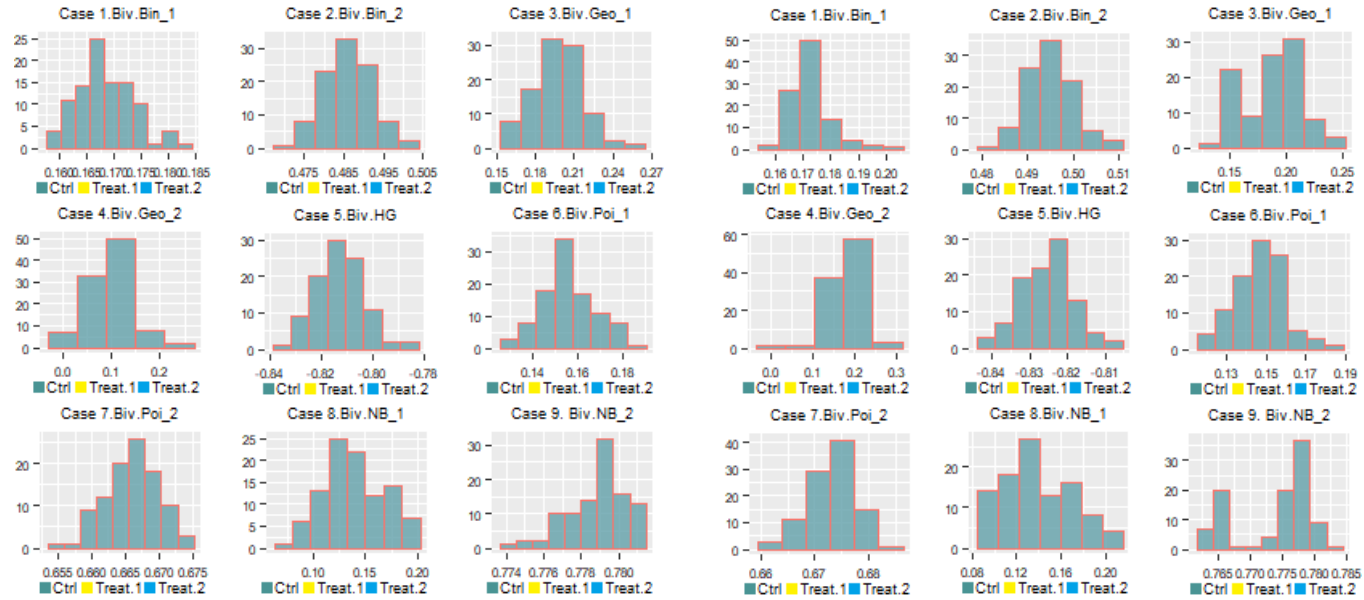

Figure 2. The frequency histograms of the τ_E (left-hand panels) and τ_C (right-hand panels) for the nine cases analyzed

4. APPLICATION TO SEISMICITY DATA

4.1 Research area and seismicity data

The Sumatra megathrust region of Indonesia is characterized by high seismic activity due to the presence of the subduction process of the Indo-Australian plates into the Eurasian plates with an average rate of 4 mm/year. Also, the region has three earthquake sources, namely the Sumatra subduction zone, Mentawai fault zone, and the Sumatra fault system.

In this study, the Sumatra megathrust (as the study area) is defined to be a rectangle region with latitude between 6.2° S - 5.5° N and longitude between 93.5° E - 104.5° E. Due to the presence of large earthquake sources, the seismologists have led to divide the region into five sub-regions, which do not overlap, namely Aceh-Andaman (AA), Nias-Simelue (NS), Mentawai-Siberut (MS), Mentawai-Pagai (MP), and Enggano (EO), as seen in left-hand panels of Figure 3 [11].

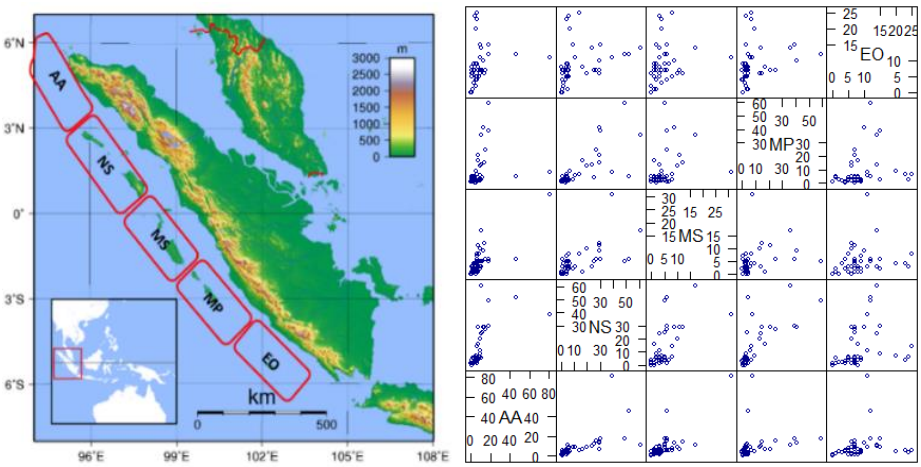

Figure 3. Map of the research area and the scatter plot of the pairs data. The left-hand panels is adopted from PusGeN [11] and Wikipedia.org (2020), while the right-hand panels is obtained using the R program

Here, the observed yearly numbers of mainshock earthquakes that occurred in the sub-regions from January 1971 to December 2018, with Magnitude of Completeness (M_c) $\geq$ 4.6 M_w are the data considered in this application.

The seismicity data source is the earthquake catalog of the United States Geological Survey (www.earthquake.usgs.gov). The data pairs of earthquakes count are graphically shown in the right-hand panels of Figure 3. Meanwhile, statistic descriptions such as min-max value, mean, and standard deviations for each sub-region are seen in Table 5.

In illustrating the proposed technique, we focus on two adjacent sub-regions of five, such as AA-NS, NS-MS, MS-MP, and MP-EO although the other six pairs are possible.

Previously, Orfanogiannaki et al. [36] and Rizal et al. [37] have provided the sub-region seismic activity in the study area using the univariate Poisson mixture models (dependent and independent). However, according to the last column in Table 4, i.e., association measure, it is important to note that for every two adjacent sub-regions, there is a positive association measure and significant (p-value < 0.01). Therefore, it shows evidence of dependence from two selected sub-regions exists. From these preliminary results, the data set can also be used in bivariate dependence modeling through the Copula model approaches.

4.2 Results

As a starting point, discrete data are transformed into continuous using the CET process proposed by Stevens [7], and then investigate the best-fit probability model among the eight models in Table 3. These models were jointly used for the investigation. In summary, those initial steps are displayed in Figure 4. The first and second rows describe the CET process, while the third and fourth describe the PDF and CDF of the selected univariate continuous model.

Table 5 shows the level comparisons of the goodness of fit (Log-llk, AIC, and BIC) between a theoretical probability model and the given data set. Here, the smallest AIC and BIC scores on each row are denoted by bold marks. However, for the sub-region MS, two criteria (AIC and BIC) gave two different models. Therefore, to overcome this problem, we use the third criterion, i.e., the maximum of Log-likelihood function (Log-llk). We then decide the appropriate model for sub-region MS would be the Gamma distribution.

Table 4. The descriptive statistics for five sub-region empirical data and association measure for two adjacent sub-regions

Sub-regions		Min	Max	Mean	Variance	Association measure	
						$\tau(i, i+1)$	*p-value*
AA	$i = 1$	0	81	8.73	165.12	0.683	< 0.01
NS	2	0	61	11.94	221.41	0.524	< 0.01
MS	3	0	31	4.58	27.56	0.285	< 0.01
MP	4	0	59	8.48	150.31	0.323	< 0.01
EO	5	0	25	8.58	32.04		

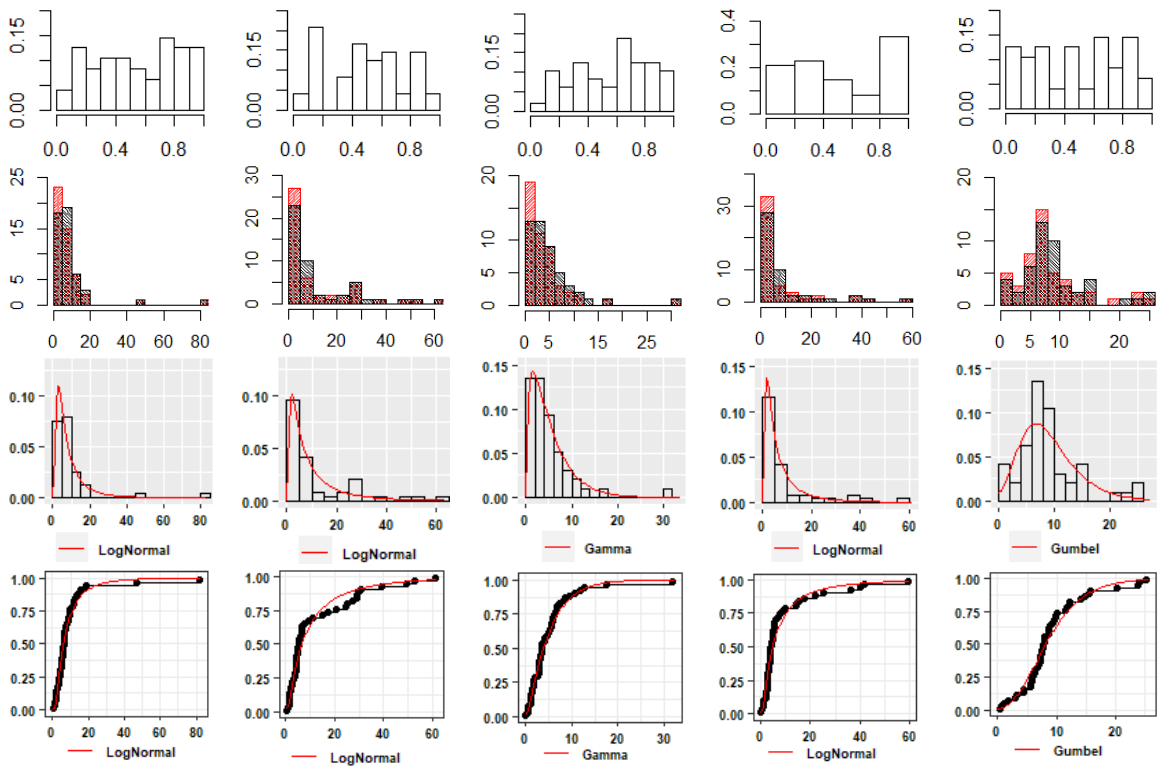

Figure 4. The visualizing of the CET process with the varying random perturbation and fitted marginal distributions. Columns correspond to the five sub-regions studied, namely AA, NS, MS, MP, and EO. Rows correspond to (1) histogram of the random perturbation (Uniform distribution), (2) histogram of the discrete variable (red color) and the continuous variable (black color), (3) density curves of fitted marginal distributions, and (4) cumulative density curve of fitted marginal distributions

With the two-information, i.e., the marginal distribution for each sub-region (Table 5), and the association measure of two adjacent sub-regions (Table 4), the next step is to search for an appropriate Copula model for the dependence structure of two selected sub-regions. To achieve this, the investigation is like the procedure of determining the marginal distribution, i.e., trial and error process. Also, among the six Copula models, each is fitted to the data set (pairs of CDFs) of two selected sub-regions. To decide the best model, the goodness-of-fit level is calculated in three terms, namely Log-likelihood, AIC, and BIC. Here, the smallest AIC and BIC scores on each row are denoted by bold marks. The investigation summary for the six Copula models and the selected model, as well as its parameter for each case studied are shown in Table 6.

Table 5. Comparison of the fitted eight continuous probability models on the basis of loglikelihood, AIC, and BIC

Sub-regions	Goodness of fit criteria	Normal	Logistik	Cauchy	Exp	LogNorm	Gamma	Weibull	Gumbel
AA	-Log-llk	190.426	171.148	154.573	154.923	147.947	153.944	154.865	160.340
	AIC	384.852	346.296	313.145	311.846	**299.895**	311.888	313.729	324.680
	BIC	388.594	350.038	316.887	313.717	**303.637**	315.630	317.472	328.422
NS	-Log-llk	196.994	193.997	180.128	168.972	165.086	168.871	168.550	183.315
	AIC	397.988	391.995	364.255	339.944	**334.172**	341.743	341.101	370.629
	BIC	401.730	395.737	367.998	341.816	**337.914**	345.485	344.843	374.372
MS	-Log-llk	147.617	138.286	135.332	126.636	125.304	**125.222**	125.879	130.082
	AIC	299.233	280.573	274.665	255.272	254.609	**254.443**	255.758	264.163
	BIC	302.975	284.315	278.407	**257.143**	258.351	258.186	259.500	267.906
MP	-Log-llk	187.842	179.539	155.090	153.410	148.818	153.345	152.850	167.278
	AIC	379.684	363.078	314.181	308.820	**301.636**	310.689	309.701	338.556
	BIC	383.427	366.820	317.923	310.691	**305.378**	314.431	313.443	342.298
EO	-Log-llk	150.461	148.092	147.702	153.799	151.791	146.167	145.532	144.900
	AIC	304.923	300.185	299.403	309.597	307.583	296.334	295.064	**293.800**
	BIC	308.665	303.927	303.146	311.469	311.325	300.077	298.807	**297.542**

Table 6. Comparison of the fitted six Copula models on the basis of AIC and BIC

Two adjacent sub-regions	Copula	Log-likelihood	AIC	BIC	Selected copula	
					Type	Parameters
AA-NS AA ~ LogNormal (1.80, 0.76) NS ~ LogNormal (1.89, 1.30)	Gaussian	26.113	-50.226	-48.354	Gumbel Hougaard	2.706
	Student's	26.073	-48.147	-44.404		
	Clayton	25.855	-49.709	-47.838		
	Gumbel	27.414	**-52.827**	**-50.956**		
	Frank	25.356	-48.711	-46.840		
	Joe	25.913	-49.826	-47.955		
NS-MS NS ~ LogNormal (1.89, 1.30) MS ~ Gamma (1.38, 0.27)	Gaussian	15.868	**-29.735**	**-27.864**	Gaussian	0.727
	Student's	15.790	-27.580	-23.837		
	Clayton	14.966	-27.932	-26.061		
	Gumbel	15.829	-29.658	-27.787		
	Frank	14.900	-27.800	-25.929		
	Joe	14.275	-26.550	-24.679		
MS-MP MS ~ Gamma (1.38, 0.27) MP ~ LogNormal (1.58, 1.24)	Gaussian	4.496	-6.993	-5.121	Clayton	0.798
	Student's	4.358	-4.717	-0.974		
	Clayton	5.738	**-9.476**	**-7.605**		
	Gumbel	4.984	-7.968	-6.097		
	Frank	4.178	-6.356	-4.485		
	Joe	5.381	-8.763	-6.891		
MP-EO MP ~ LogNormal (1.58, 1.24) EO ~ Gumbel (6.61, 4.21)	Gaussian	5.652	**-9.303**	**-7.432**	Gaussian	0.501
	Student's	5.480	-6.960	-3.217		
	Clayton	4.816	-7.632	-5.761		
	Gumbel	5.026	-8.051	-6.180		
	Frank	5.284	-8.568	-6.697		
	Joe	4.183	-6.366	-4.495		

4.3 Comparison with other count models

In this section, we compare our proposed technique, i.e., Copula continuous extension technique, and the popular models for count data, such as bivariate Negative Binomial [38, 39], bivariate Poisson [40, 41], and double Poisson distribution [12]. The first two models have been briefly described in subsection 2.1, as seen in Table 1.

The double Poisson is a special case of the Poisson bivariate. It is important to consider the bivariate Poisson formula for bivariate discrete (X, Y), denoted by $\text{BivPoi}(\lambda_1, \lambda_2, \lambda_3)$. Marginally, each random variable follows a univariate Poisson distribution with $E[X] = \text{Var}[X] = \lambda_1 + \lambda_3$, $E[Y] = \text{Var}[Y] = \lambda_2 + \lambda_3$, and $\text{Cov}[X, Y] = \lambda_3$. The parameter λ_3 is a measure of dependence between the two random variables X and Y. If $\lambda_3 = 0$, then the two variables are independent and the bivariate Poisson formula is represented as the product of two independent univariate Poisson distributions (called as double Poisson distribution).

The comparison of four models selected for count data

modeling in Table 7 shows that, on the basis of AIC and BIC, the Copula continuous extension technique is superior to the models of double Poisson, bivariate Poisson, and bivariate Negative Binomial.

Some aspects can be read from Table 7 which is associated with the information in Table 4. Firstly, it is reasonable that double Poisson distribution has the last rank, since the data set of the two selected sub-regions studied were not independent (Table 4). Secondly, it is important to note that according to the fifth and sixth columns of Table 4, an overdispersion condition, i.e., variance greater than mean, in the observation data from each sub-region exists. As we know, the marginal of the bivariate Negative Binomial allows for overdispersion found in the data set, which can not be accounted by the bivariate Poisson. Therefore, the bivariate Negative Binomial have a better performance than the bivariate Poisson. Thirdly, although the bivariate Negative Binomial can overcome the problems of dependent and overdispersion in the data set, it is not enough to capture the joint probability behavior of our bivariate data, as well as the selected Copula model, due to this model have assumptions that must be considered [12, 13].

Finally, the joint probability function is determined from the two adjacent sub-regions considered by combining the selected marginals (Table 5) and selected the Copula model (Table 6). For two adjacent sub-regions AA-NS, NS-MS, MS-MP, and MP-EO are expressed as follows, respectively.

$$H_{\mathrm{AA,NS}}(x,y)=\exp\left(-\left(\left(-\log\left(\frac{1}{2}\left(1+\operatorname{erf}\left(\frac{\ln x-1.80}{1.232}\right)\right)\right)\right)^{2.706}+\left(-\log\left(\frac{1}{2}\left(1+\operatorname{erf}\left(\frac{\ln y-1.89}{1.612}\right)\right)\right)\right)^{2.706}\right)^{0.370}\right) \tag{12}$$

$$H_{\mathrm{NS,MS}}(x,y)=\Phi_G\left[\Phi^{-1}\left(\frac{1}{2}\left(1+\operatorname{erf}\left(\frac{\ln x-1.80}{1.232}\right)\right)\right),\Phi^{-1}\left(\frac{1}{\Gamma(1.38)}\gamma\left(1.38,\frac{x}{0.27}\right)\right);0.727\right] \tag{13}$$

$$H_{\mathrm{MS,MP}}(x,y)=\left(\left[\frac{1}{\Gamma(1.38)}\gamma\left(1.38,\frac{x}{0.27}\right)\right]^{-0.798}+\left[\frac{1}{2}\left(1+\operatorname{erf}\left(\frac{\ln x-1.58}{1.575}\right)\right)\right]^{-0.798}-1\right)^{-1.253} \tag{14}$$

$$H_{\mathrm{MP,EO}}(x,y)=\Phi_G\left[\Phi^{-1}\left(\frac{1}{2}\left(1+\operatorname{erf}\left(\frac{\ln x-1.58}{1.575}\right)\right)\right),\Phi^{-1}\left(\exp\left(-e^{-\frac{x-6.61}{4.21}}\right)\right);0.501\right] \tag{15}$$

Another way of visualizing the selected Copula model is to present its densities as contour plots (first row), as shown in Figure 5. Furthermore, as seen in the second row in Figure 5, the scatter plots of simulated and observed data are shown. Also, the association measure of the simulated data of AA-NS, NS-MS, MS-MP, and MP-EO were 0.668, 0.514, 0.273, and 0.314, respectively. These results are near to the association measures from the observed data, i.e., 0.683, 0.524, 0.285, and 0.323, as shown in Table 4. In other words, the selected Copula models yielding from the Copula continuous extension technique can capture the association behavior of bivariate discrete variables.

Table 7. Comparison of the four models based on AIC and BIC

Two adjacent sub-regions	Bivariate Poisson		Double Poisson		Bivariate Negative Binomial		Copula continuous extension technique		
	AIC	BIC	AIC	BIC	AIC	BIC	Model	AIC	BIC
AA-NS	1341.643	1349.336	1516.381	1521.510	330.843	336.456	Gumbel	-52.827	-50.956
NS-MS	1137.008	1144.701	1223.226	1228.355	373.964	379.577	Gaussian	-29.735	-27.864
MS-MP	1035.361	1043.055	1059.908	1065.037	296.043	307.656	Clayton	-9.476	-7.605
MP-EO	1058.489	1066.182	1067.261	1072.389	319.571	325.184	Gaussian	-9.303	-7.432

As a note, the AIC-BIC listed in the Copula continuous extension technique refers to the selected Copula model as shown in Table 6.

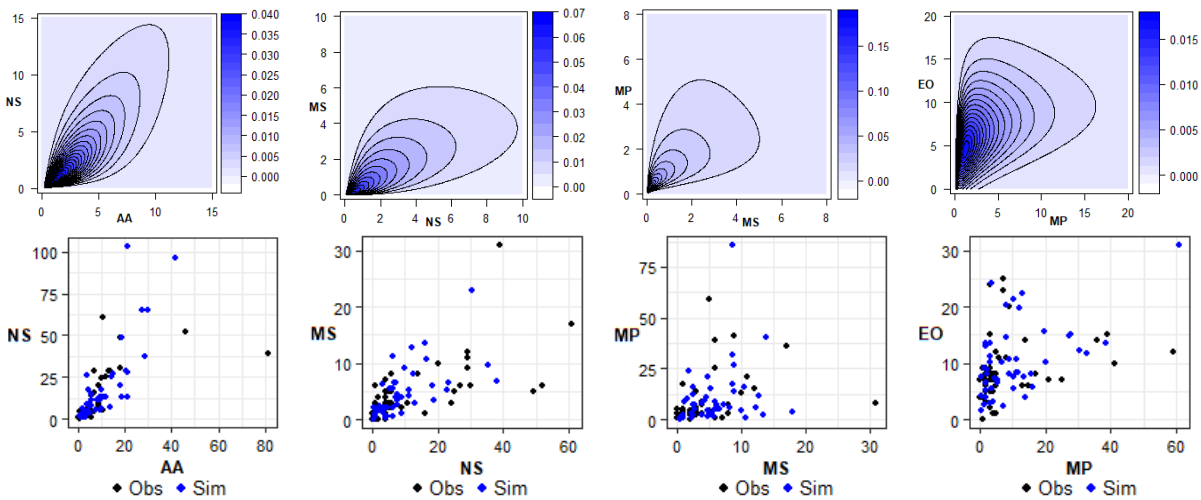

Figure 5. Some visualizations are based on selected Copula models. Columns correspond to sub-regions studied, namely AA-NS, NS-MS, MS-MP, and MP-EO. Rows correspond to contour plots of the densities, and scatter plots observed and simulated

5. CONCLUSIONS

A simple technique has been provided in modifying the Copula modeling method for bivariate discrete variables, which is the continuous extension. There are two steps carried out in this technique. The first is applying the Continuous Extension Technique (CET) at the marginal bivariate discrete variables to become continuous ones. In the second step, a Copula model was constructed for a new bivariate variable (continuous) by following the procedure from the Copula theory for continuous variables. However, two forms of random perturbation of the CET can be used, i.e., $U(0,1)$ and $(U(0,1)-1)$ as proposed by Stevens [7] and modified by Denuit and Lambert [9], respectively. Based on analytical and simulation studies, two different forms of CET produced the same conclusions about Kendall's Tau measure and the selected Copula model. Therefore, practitioners can select one of two existing CET forms in practice.

The advantage of the present technique is not only from practical point of view, but it is also able to preserve the parameter dependence of the former discrete variables. Thus, the selected Copula model based on the Copula continuous extension technique can be used to represent the dependence modeling of the discrete variables.

To help the seismologists understand the proposed technique, an application to the seismicity data recorded in five sub-regions of the Sumatra megathrust, namely Aceh-Andaman (AA), Nias-Simelue (NS), Mentawai-Siberut (MS), Mentawai-Pagai (MP), and Enggano (EO) has been presented. The illustration was based on the observed yearly numbers of mainshock earthquakes that occurred from January 1971 to December 2018, with Magnitude of Completeness (M_c) 4.6 M_w. According to the selected Copula model, the evidence of dependence seismic activity in each of the two adjacent sub-regions exists. The results of these analyses provide new information regarding the seismicity behavior in the Sumatra megathrust. Therefore, as future study, the dependence modeling using the Copula models can be applied as an alternative approach in developing the prediction of earthquake hazard models in Sumatra megathrust.

ACKNOWLEDGMENT

The authors sincerely wish to thank the journal editor and reviewers who critically reviewed the manuscript and made valuable suggestions for its improvement. The first author would like to thank LPDP (Lembaga Pengelola Dana Pendidikan), Ministry of Finance, Republic Indonesia for providing the Ph.D. scholarship program. Partial funding is also supported by P2MI research grant of Institut Teknologi Bandung, Indonesia 2021 (Grant numbers: 62/IT1.C02/SK-TA/2021).

REFERENCES

[1] Zhang, X., Jiang, H. (2019). Application of Copula function in financial risk analysis. Computers & Electrical Engineering, 77: 376-388. https://doi.org/10.1016/j.compeleceng.2019.06.011

[2] Bhatti, M.I., Do, H.Q. (2019). Recent development in copula and its applications to the energy, forestry and environmental sciences. International Journal of Hydrogen Energy, 44(36): 19453-19473. https://doi.org/10.1016/j.ijhydene.2019.06.015

[3] Dong, H., Xu, X., Sui, H., Xu, F., Liu, J. (2017). Copula-based joint statistical model for polarimetric features and its application in PolSAR image classification. IEEE Transactions on Geoscience and Remote Sensing, 55(10): 5777-5789. https://doi.org/10.1109/TGRS.2017.2714169

[4] Ghahroodi, Z.R., Saba, R.A., Baghfalaki, T. (2019). Gaussian copula-based regression models for the analysis of mixed outcomes: An application on household's utilization of health services data. Journal of Statistical Theory and Applications, 18(3): 182-197. https://doi.org/10.2991/jsta.d.190306.009

[5] Wan, J., Li, S. (2019). Modeling and application of industrial process fault detection based on pruning vine copula. Chemometrics and Intelligent Laboratory Systems, 184: 1-13. https://doi.org/10.1016/j.chemolab.2018.11.005

[6] Marshall, A. (1996). Copulas, marginals, and joint distributions. Institute of Mathematic Statistics, 28: 213-222. https://doi.org/10.1214/lnms/1215452620

[7] Stevens, W.L. (1950). Fiducial limits of the parameter of a discontinuous distribution. Biometrika, 37(1/2): 117-129. https://doi.org/10.1093/biomet/37.1-2.117

[8] Machado, J.A.F., Santos Silva, J.M.C. (2005). Quantiles for counts. Journal of the American Statistical Association, 100(472): 1226-1237. https://doi.org/10.1198/016214505000000330

[9] Denuit, M., 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

[10] 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.

[11] PusGeN (2017). Peta Sumber dan Bahaya Gempa Indonesia Tahun 2017, Pusat Penelitian dan Pengembangan Perumahan dan Pemukiman. Badan Peneliti dan Pengembangan Kementerian Pekerjaan Umum dan Perumahan Rakyat, pp. 89-91 and 189 (in indonesian). Documentation available at https://www.scribd.com/document/394240076/BUKU-PETA-GEMPA-2017-pdf.

[12] Kocherlakota, S., Kocherlakota, K.K. (1992). Bivariate Discrete Distributions. Marcel Decker, New York. https://archive.org/details/bivariatediscret00koch.

[13] Montgomery, D.C., Runger, G.C. (2007). Applied Statistics and Probability for Engineers. John Wiley & Sons.

[14] Joe, H. (1997). Multivariate Models and Multivariate Dependence Concepts. CRC Press. https://doi.org/10.1201/9780367803896

[15] Nelsen, R.B. (2006), An Introduction to Copulas. 2nd edition. New York: Springer. https://doi.org/10.1007/0-387-28678-0

[16] Hofert, M., Kojadinovic, I., Mächler, M., Yan, J. (2019). Elements of Copula Modeling with R. Springer. https://doi.org/10.1007/978-3-319-89635-9

[17] Marshall, A.W., Olkin, I. (1985). A family of bivariate distributions generated by the bivariate Bernoulli distribution. Journal of the American Statistical Association, 80(390): 332-338. https://doi.org/10.2307/2287890

[18] Trivedi, P., 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
[19] Kruskal, W.H. (1958). Ordinal measures of association. Journal of the American Statistical Association, 53(284): 814-861. https://doi.org/10.1080/01621459.1958.10501481
[20] Lehmann, E.L. (1966). Some concepts of dependence. The Annals of Mathematical Statistics, 37(5): 1137-1153. https://doi.org/10.1214/aoms/1177699260
[21] Agresti, A. (2010). Analysis of Ordinal Categorical Data (Second ed.). New York: John Wiley & Sons. https://doi.org/10.1002/9780470594001
[22] Cherubini, U., Luciano, E., Vecchiato, W. (2004). Copula Methods in Finance West Sussex: John Wiley and Sons. https://doi.org/10.1002/9781118673331
[23] Angus, J.E. (1994). The probability integral transform and related results. SIAM Review, 36(4): 652-654. https://doi.org/10.1137/1036146
[24] Joe, H., Xu, J.J. (1996). The estimation method of inference functions for margins for multivariate models. Technical Report 166, Department of Statistics, University of British Columbia. https://dx.doi.org/10.14288/1.0225985
[25] Ross, S.M. (2014). Introduction to Probability Models. Academic Press. https://doi.org/10.1016/C2012-0-03564-8
[26] Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6): 716-723. https://doi.org/10.1109/TAC.1974.1100705
[27] Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2): 461-464. https://doi.org/10.1214/aos/1176344136
[28] Wolodzko, T. (2019). ExtraDistr: Additional Univariate and Multivariate Distributions,. R Package Version, 1(11). https://cran.r-project.org/web/packages/extraDistr/extraDistr.pdf.
[29] de Oliveira, R.P., Achcar, J.A. (2018). Basu-Dhar's bivariate geometric distribution in presence of censored data and covariates: Some computational aspects. Electronic Journal of Applied Statistical Analysis, 11(1): 108-136. https://doi.org/10.1285/i20705948v11n1p108
[30] Spurdle, A. (2020). Bivariate (Version 0.6.0). https://cran.r-project.org/web/packages/bivariate/index.html.
[31] Delignette-Muller, M.L., Dutang, C. (2015). Fitdistrplus: An R package for fitting distributions. Journal of Statistical Software, 64(4): 1-34. https://doi.org/10.18637/jss.v064.i04
[32] Kirkpatrick, R.M. (2014). RMKdiscrete (Version 0.1). http://cran.r-project. org/web/packages/RMKdiscrete.
[33] Hofert, M., Kojadinovic, I., Maechler, M., Yan, J., Maechler, M.M., Suggests, M.A.S.S. (2014). Package 'copula'. http://ie.archive.ubuntu.com/disk1/disk1/cran.r-project.org/web/packages/copula/copula. pdf.
[34] Schepsmeier, U., Stoeber, J., Brechmann, E.C., Graeler, B., Nagler, T., Erhardt, T., Killiches, M. (2015). Package 'VineCopula'. R package version, 2(5). https://cran.r-project.org/web/packages/VineCopula/index.html.
[35] Xu, Y., Tang, X.S., Wang, J.P., Kuo-Chen, H. (2016). Copula-based joint probability function for PGA and CAV: A case study from Taiwan. Earthquake Engineering and Structural Dynamics, 45(13): 2123-2136. https://doi.org/10.1002/eqe.2748
[36] 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). https://doi.org/10.4081/rg.2014.4969
[37] 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, 1097(1): 012083. https://iopscience.iop.org/article/10.1088/1742-6596/1097/1/012083.
[38] Iwasaki, M., Tsubaki, H. (2006). Bivariate negative binomial generalized linear models for environmental count data. Journal of Applied Statistics, 33(9): 909-923. https://doi.org/10.1080/02664760600744157
[39] Fernando, S.M., Sooriyarachchi, M.R. (2018). Bivariate negative binomial modelling of epidemiological data. Open Science Journal of Statistics and Application, 5(3): 47.
[40] Karlis, D., Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. Journal of the Royal Statistical Society: Series D (The Statistician), 52(3): 381-393. https://doi.org/10.1111/1467-9884.00366
[41] Berkhout, P., Plug, E. (2004). A bivariate Poisson count data model using conditional probabilities. Statistica Neerlandica, 58(3): 349-364. https://doi.org/10.1111/j.1467-9574.2004.00126.x

NOMENCLATURE

CET	Continous extension technique
PCC	Pearson correlation coefficient
U(0,1)	Uniform distribution (0,1)
V(0,1)	Uniform distribution (0,1) that independent with U(0,1)
X	Discrete random variable
$X^{\star}$	A continuous random variable that resulting from the CET process with the Stevens method.
$X^{\circledast}$	A continuous random variable that resulting from the CET process with the Denuit and Lambert method.
E []	The expectation of arandom variable
Var ()	The variance of a random variable
Cov ()	The covariance of the two variables
F()	The cumulative distribution function
PDF	Probablitiy distribution function
PMF	Probablitiy mass function
CDF	Cumulative distribution function
$H()$	Joint cumulative distribution
$C()$	The Copula model
erf (x)	$\frac{2}{\sqrt{\pi}}\int_0^x e^{t^2}dt.$
T	The number of observation
R	Real numbers
p-value	the probability of declaring that the test result is at least the same as the actual observed result, assuming that the null hypothesis is correct.
log	Logarithm function
Log-llk	Logarithm of the loglikelihood function
n/a	not available
AIC	Akaike Information Criteria

BIC	Bayesian Information Criteria
$\mathrm{p}(l)$	the number of parameters of the fitted model
$\mathrm{M_c}$	Magnitude of Completeness
$\mathrm{M_w}$	moment magnitude
$\subseteq$	Subset

Greek symbols

τ	Kendall's Tau measure
ρ	Pearson correlation coefficient
σ	Standard deviation
$\mathcal{L}$	The likelihood function
θ	Parameter of Copula models
$\boldsymbol{\omega}$	Parameters of univariate continuous models
Φ	CDF of a standard normal distribution
Φ^{-1}	Invers of Φ
ν	Degrees of freedom of the bivariate t-distribution
Γ	Gamma function