



Vine Copula Model: Application Chemical Elements in Water Samples

S Z Aminullah^{1,*}, M Novita¹, I Fithriani¹

¹ Department of Mathematics, Universitas Indonesia, Depok, 16424, Indonesia

*Corresponding author's e-mail: salsabila.zahra@sci.ui.ac.id

Abstract. Copula can link the bivariate distribution function with marginal distribution functions without requiring specific information about the interdependence among random variables. There are several types of copulas, such as elliptical copulas, Archimedean copulas, and extreme value copulas. However, in multivariate modeling, each type of copula has limitations in modeling complex dependence structures in terms of symmetry and tail dependence properties. The class of vine copulas overcomes these limitations by constructing multivariate models using bivariate copulas in a tree-like structure. The bivariate copulas used in this study include the Clayton, Gumbel, Frank, Gaussian, and Student's t copula families. This study discusses the construction of vine copula models, parameter estimation, and their applications. The construction of vine copulas is done through the decomposition of conditional probability density functions and substituting bivariate copula density functions into the decomposition results. The data used in the study is the logarithm of the concentration of chemical elements in water samples in Colorado. The parameter estimation method used is pseudo-maximum likelihood with sequential estimation. Model selection is then performed using the Akaike information criterion (AIC) to determine the most suitable model. The results indicate that Caesium and Titanium have a dependency relationship with Scandium. Moreover, Scandium and Titanium exhibit the strongest dependence compared to other variable pairs.

1. Introduction

Each event in life occurs due to various causes and interacting conditions that are interconnected. Everything that exists is a condition that influences others, and vice versa. There is a complex interrelation among various factors and conditions that mutually influence each other [10]. An event can be influenced by multiple factors simultaneously, and its impact can also spread to other factors and conditions. For instance, in an economic context, if a country drastically raises its interest rates, it can increase investor interest in investing their money in that country's currency. The demand for that currency will increase, potentially leading to an appreciation in the currency's exchange rate, and vice versa. To measure this interdependence, it can be examined through the covariance of these two factors. Although covariance provides information about the nature of the relationship between two random variables, it does not indicate the strength of the relationship because covariance is not scale-free; its magnitude depends on the units used to measure both variables [30]. There is a scale-free version of covariance known as the correlation coefficient. The most common form of this linear correlation coefficient is known as Pearson's correlation coefficient. According to Czado [8], Pearson's correlation coefficient is a measure of linear dependence whose values range between -1 and 1.



Despite Pearson's correlation coefficient being widely used to measure dependence between variables, there are limitations to its application. One limitation is that it cannot explain the relationship between randomly distributed non-elliptical variables [11]. An elliptical distribution is an extension of the multivariate normal distribution, where the distribution's shape can be represented by an ellipsoid.

To address the limitations of Pearson's correlation coefficient, the use of copulas can be considered. According to Nelsen [24], the term "copula" originates from Latin and means "tie, bond, or link" (Latin Dictionary Cassell). The concept of copulas was first introduced in the mathematical and statistical context by Abe Sklar in 1959 in Sklar's theorem. Based on this theorem, a copula is a function that connects a multivariate distribution function with its marginal cumulative distribution functions. The idea of copulas had previously appeared in various articles without referring to them as copulas, e.g., Hoeffding [15][16].

There are three main approaches to copula construction as outlined by Czado [8]. The first approach involves applying probability integral transformations to the marginal distributions of a known multivariate distribution. This approach is applied to elliptical distributions, resulting in the class of elliptical copulas, including Gaussian copulas and Student's t copulas. The second approach employs generator functions, resulting in the class of Archimedean copulas, which includes the Clayton, Gumbel, Frank, and Joe copula families. The third approach extends the univariate extreme value theory to higher dimensions, resulting in the class of extreme value copulas, including the Marshall–Olkin copula and the Hüsler–Reiss copula.

The copulas from these three classes are less flexible in higher dimensions because they cannot define different tail dependencies for different variable pairs. Moreover, the limitation of Archimedean copulas lies in the fact that their dependence structure is determined by a single parameter, limiting their flexibility in modeling complex dependence structures [5]. Elliptical copulas have limitations in modeling asymmetric dependence [26]. Extreme value copulas can only model one type of tail dependence [2]. The vine copula model addresses these issues by combining bivariate copulas in a tree structure, allowing for the modeling of complex dependence structures.

The vine copula model was initially proposed by Joe [21] and further developed in Bedford and Cooke [3][4]. The vine copula employs bivariate copulas to decompose multivariate distribution functions (with more than two variables). Bedford and Cooke introduced a hierarchical tree structure to organize the decomposition, naming the resulting graphical structure a "vine." In this tree structure, nodes represent marginal density functions and bivariate copulas, while edges represent bivariate copulas used to connect these nodes.

The primary advantage of the vine copula model is that all involved copulas are bivariate, and they do not need to be identical for all variable pairs. Furthermore, each copula pair has its parameter to be estimated. In this study, the parameter estimation will be performed using the pseudo-maximum likelihood method. Pseudo-maximum likelihood is similar to the maximum likelihood method; it estimates parameters by maximizing the likelihood function of observed data. The difference lies in the data used to maximize the likelihood function. In pseudo-maximum likelihood, the marginal distribution is unknown. In vine copulas, sequential estimation is performed, estimating from the first tree to the last [8].

The vine copula model can be applied in various fields, such as the financial industry, environmental science, health, and technology. Nikoloulopoulos [35] evaluated the accuracy of diagnostic tests using a trivariate vine copula model for true positive (correctly diagnosed sick individuals), true negative (correctly diagnosed healthy individuals), and sick individuals. Hohndorf [18] analyzed relationships between variables emerging from operational flight data using marginal regression with vine copulas. Additionally, Ernhardt [12] used vine copulas to model various types of health insurance claims over multiple time periods. In this research, the construction of the vine copula model and its utilization in assessing the dependence pattern among various chemical elements in water samples will be outlined. Subsequently, model selection will be performed based on the Akaike information criterion.



2. Copula

A copula is a multivariate distribution function defined within the $[0, 1]^n$ hypercube, where each variable follows a uniform distribution. In [24], according to Sklar's Theorem, let H be an n -dimensional distribution function with marginal distributions F_1, F_2, \dots, F_n . Then, there exists an n -copula C such that for all x in \bar{R}^n :

$$H(x_1, x_2, \dots, x_n) = C(F_1(x_1), F_2(x_2), \dots, F_n(x_n)) \quad (1)$$

The probability density function of (1) is as follows:

$$\begin{aligned} h(x_1, x_2, \dots, x_n) &= \frac{\partial^{(n)}}{\partial x_1 \partial x_2 \dots \partial x_n} H(x_1, x_2, \dots, x_n) \\ &= c(F_1(x_1), F_2(x_2), \dots, F_n(x_n)) f_n(x_n) f_{n-1}(x_{n-1}) \dots f_1(x_1) \end{aligned} \quad (2)$$

Let $F_1^{-1}, F_2^{-1}, \dots, F_n^{-1}$ be the inverses of F_1, F_2, \dots, F_n respectively. Then, for every u in I^n

$$C(u_1, u_2, \dots, u_n) = H(F_1^{-1}(u_1), F_2^{-1}(u_2), \dots, F_n^{-1}(u_n)) \quad (3)$$

According to Czado [8], the copula density function can be obtained through partial differentiation, as follows

$$\begin{aligned} c(u_1, u_2, \dots, u_n) &= \frac{\partial^{(n)}}{\partial u_1 \partial u_2 \dots \partial u_n} H(F_1^{-1}(u_1), F_2^{-1}(u_2), \dots, F_n^{-1}(u_n)) \\ &= \frac{h(F_1^{-1}(u_1), F_2^{-1}(u_2), \dots, F_n^{-1}(u_n))}{h_n(F_1^{-1}(u_n)) h_{n-1}(F_2^{-1}(u_{n-1})) \dots h_1(F_1^{-1}(u_1))} \end{aligned} \quad (4)$$

3. Tail Dependence

According to Nelsen [24], the tail dependence measures the dependence between variables in the upper-right and lower-left tails of a bivariate distribution. Let X_1 and X_2 be continuous random variables with distribution functions F_1 and F_2 , respectively. The upper tail dependence parameter λ_{upper} is defined as follows:

$$\lambda_{upper} = \lim_{t \rightarrow 1^-} P(X_2 > F_2^{-1}(t) | X_1 > F_1^{-1}(t)) = \lim_{t \rightarrow 1^-} \frac{1 - 2t + C(t, t)}{1 - t} \quad (5)$$

The lower tail dependence parameter λ_{lower} is defined as follows:

$$\lambda_{lower} = \lim_{t \rightarrow 0^+} P(X_2 \leq F_2^{-1}(t) | X_1 \leq F_1^{-1}(t)) = \lim_{t \rightarrow 0^+} \frac{C(t, t)}{t} \quad (6)$$

4. Bivariate Copula Family

In this research, a bivariate copula, also known as a two-dimensional copula, is the most commonly used copula in modeling the dependence between random variables. In this study, the Elliptical copula class is used, which includes the Gaussian copula and the Student's t copula, as well as the Archimedean copula class, which includes the Gumbel copula, the Clayton copula, and the Frank copula.

4.1 Elliptical Copula

Elliptical copulas are constructed from elliptical distributions using Sklar's theorem by applying probability integral transformations to each marginal distribution of the known multivariate elliptical distribution.

4.1.1 Gaussian Copula. A bivariate Gaussian copula is obtained by using a bivariate normal distribution with a zero mean vector, unit variance, and correlation ρ . Applying the inverse of Sklar's theorem, the cumulative distribution function of (U_1, U_2) is as follows

$$C(u_1, u_2; \rho) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho) \quad (7)$$



where $(u_1, u_2) \in [0,1]^2$, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution $N(0,1)$ and $\Phi_2(\cdot, \cdot; \rho)$ is the bivariate normal cumulative distribution function with a zero mean, unit variance, and correlation ρ [8]. The Gaussian copula has only one parameter ρ . As for the copula density function, it is as follows

$$c(u_1, u_2, \rho) = \frac{1}{\sqrt{1-\rho^2}} \exp \left\{ -\frac{\rho^2(x_1^2 + x_2^2) - 2\rho x_1 x_2}{2(1-\rho)} \right\} \quad (8)$$

4.1.2 Student's t Copula. The Student's t copula can be constructed using the Student's t distribution with degrees of freedom ν , a zero mean, and correlation ρ . Sklar's theorem yields its cumulative distribution function as follows

$$C(u_1, u_2; \nu, R) = T_{\nu, R}(T_{\nu}^{-1}(u_1), T_{\nu}^{-1}(u_2)) \quad (9)$$

$T_{\nu, R}$ represents the cumulative distribution function of the standard bivariate Student's t distribution with $\nu > 0$. T_{ν}^{-1} is the inverse of the cumulative distribution function T_{ν} which is the standard univariate cumulative distribution function of the Student's t distribution with degrees of freedom ν [8]. The copula density function is as follows

$$c(u_1, u_2; \nu, R) = \frac{\frac{1}{2\pi\sqrt{1-\rho^2}} \left[1 + \frac{1}{\nu(1-\rho^2)} (x_1 - 2\rho x_1 x_2 + x_2^2) \right]^{-\frac{\nu+2}{2}}}{\left(\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})} \right)^2 \left[\left(1 + \frac{x_1^2}{\nu} \right) \left(1 + \frac{x_2^2}{\nu} \right) \right]^{-\frac{\nu+1}{2}}} \quad (10)$$

4.2 Archimedean Copula

Within this section, Archimedean copulas will be. The class of Archimedean copulas allows for various different dependence structures. An Archimedean copula is constructed from a generator function. Let ψ be a continuous function that is strictly decreasing from $[0,1]$ to $[0, \infty]$ where $\psi(1) = 0$, and $\psi^{[-1]}$ is the pseudo-inverse of ψ . Let C be a function from $[0,1]^2$ to $[0,1]$ given by

$$C(u_1, u_2) = \psi^{[-1]}(\psi(u_1) + \psi(u_2)) \quad (11)$$

Then C is a copula if and only if ψ is a convex function [11].

4.2.1 Gumbel Copula. The Gumbel Copula is characterized by a single parameter $\theta \geq 1$. When $\theta = 1$, the Gumbel Copula represents independence. As $\theta \rightarrow \infty$, it indicates complete dependence. The Gumbel Copula has the following generator function:

$$\psi_{\theta}(t) = (-\ln t)^{\theta} \quad (12)$$

The cumulative distribution function of the Gumbel Copula is given by

$$C_{\theta}(u_1, u_2) = \exp \left(-\left((-\ln u_1)^{\theta} + (-\ln u_2)^{\theta} \right)^{\frac{1}{\theta}} \right) \quad [8] \quad (13)$$

The density function for the Gumbel Copula is as follows

$$c_{\theta}(u_1, u_2) = \frac{((-\ln u_1)(-\ln u_2))^{\theta-1}}{u_1 u_2} \left((\theta-1) \left((-\ln u_1)^{\theta} + (-\ln u_2)^{\theta} \right)^{\frac{1}{\theta}-2} + \left((-\ln u_1)^{\theta} + (-\ln u_2)^{\theta} \right)^{\frac{2}{\theta}-2} \right) \quad (14)$$

4.2.2 Clayton Copula. The Clayton Copula has a single parameter $\theta > 0$. As $\theta \rightarrow 0$, the Clayton Copula represents no dependence, and as $\theta \rightarrow \infty$, it indicates complete dependence. The Clayton Copula has the following generator function



$$\psi_{\theta}(t) = \frac{1}{\theta}(t^{-\theta} - 1) \quad (15)$$

The cumulative distribution function of the Clayton Copula is given by

$$C_{\theta}(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-\frac{1}{\theta}} \quad [8] \quad (16)$$

The density function for the Clayton Copula is as follows

$$c(u_1, u_2) = (\theta + 1)(u_1 u_2)^{-\theta-1}(u_1^{-\theta} + u_2^{-\theta} - 1)^{-\frac{1}{\theta}-2} \quad (17)$$

4.2.3 *Frank Copula*. The Frank Copula has a parameter $\theta \in [-\infty, \infty] \setminus \{0\}$. As $\theta \rightarrow 0$, it represents independence. The Frank Copula has the following generator function

$$\psi_{\theta}(t) = -\ln\left(\frac{e^{-\theta t} - 1}{e^{-\theta} - 1}\right) \quad (18)$$

The cumulative distribution function of the Frank Copula is given by

$$C_{\theta}(u_1, u_2) = -\frac{1}{\theta} \ln\left(1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{(e^{-\theta} - 1)}\right) \quad (19)$$

$$C_{\theta}(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-\frac{1}{\theta}} \quad [8] \quad (20)$$

The cumulative distribution function of the Frank Copula is given by

$$c(u_1, u_2) = \frac{\theta e^{-\theta(u_1+u_2)}(e^{-\theta} - 1)}{(e^{-\theta(u_1+u_2)} - e^{-\theta u_1} - e^{-\theta u_2} + e^{-\theta})^2} \quad (21)$$

5. Regular Vine Copula Model

5.1. Within this section, an explanation of pair copula construction and a regular vine in three dimensions will be provided. *Pair-Copula Construction (PCC)*

The Pair-Copula Construction (PCC) was initially proposed by Joe [21] and further developed by Bedford & Cooke [3][4]. The fundamental idea behind PCC is to construct higher-dimensional copulas through bivariate copulas, which provide a flexible class of dependence models. To illustrate the concept of Pair-Copula Construction, it is necessary to introduce the decomposition of pair copulas from the multivariate density function. Let X_1, \dots, X_n be random variables. The probability density function for the multivariate case can be expressed as a series of conditional univariate density functions. In the case of two variables, the joint density function can be expressed as

$$f(x_1, x_2) = f(x_2|x_1)f_1(x_1) \quad (22)$$

For three variables, one of the forms is

$$f(x_3|x_1, x_2) = \frac{f(x_1, x_2, x_3)}{f(x_1, x_2)} \quad (23)$$

$$f(x_1, x_2, x_3) = f(x_3|x_1, x_2)f(x_1, x_2) \quad (24)$$

For four variables, one of the forms is

$$f(x_4|x_1, x_2, x_3) = \frac{f(x_1, x_2, x_3, x_4)}{f(x_1, x_2, x_3)} \quad (25)$$

$$f(x_1, x_2, x_3, x_4) = f(x_4|x_1, x_2, x_3)f(x_1, x_2, x_3) \quad (26)$$

This pattern continues, and for n variables, one of the forms is

$$f(x_n|x_1, \dots, x_{n-1}) = \frac{f(x_1, \dots, x_n)}{f(x_1, \dots, x_{n-1})} \quad (27)$$

$$f(x_1, \dots, x_n) = f(x_n|x_1, \dots, x_{n-1})f(x_1, \dots, x_{n-1}) \quad (28)$$

Thus, one of the decomposition forms is as follows



$$f(x_1, \dots, x_n) = f(x_n|x_1, \dots, x_{n-1})f(x_{n-1}|x_1, \dots, x_{n-2}) \dots f(x_2|x_1)f(x_1) \quad (29)$$

where $f(\cdot | \cdot)$ represents the conditional density function.

Based on Sklar's theorem, the density function for the bivariate case can be expressed as follows

$$f(x_1, x_2) = c_{1,2}(F_1(x_1), F_2(x_2))f_1(x_1)f_2(x_2) \quad (30)$$

Therefore, the conditional density function can be expressed as

$$f(x_2|x_1) = c_{1,2}(F_1(x_1), F_2(x_2))f_2(x_2) \quad (31)$$

For three random variables X_1, X_2 , and X_3 , the conditional density function $f(x_3|x_1, x_2)$ can also be expressed as

$$f(x_3|x_1, x_2) = \frac{f(x_1, x_2, x_3)}{f(x_1, x_2)} = \frac{f(x_2, x_3|x_1)f_1(x_1)}{f(x_2|x_1)f_1(x_1)} = \frac{f(x_2, x_3|x_1)}{f(x_2|x_1)} \quad (32)$$

According to Sklar's theorem, the conditional density function $f(x_2, x_3|x_1)$ can also be written as follows

$$f(x_2, x_3|x_1) = c_{2,3|1}(F(x_2|x_1), F(x_3|x_1)|x_1)f(x_2|x_1)f(x_3|x_1) \quad (33)$$

Therefore, it can be concluded that

$$f(x_3|x_1, x_2) = c_{2,3|1}(F(x_2|x_1), F(x_3|x_1)|x_1)f(x_3|x_1) \quad (34)$$

$$= c_{2,3|1}(F(x_2|x_1), F(x_3|x_1)|x_1)c_{1,3}(F_1(x_1), F_3(x_3))f_3(x_3) \quad (35)$$

Where $f(x_3|x_1) = c_{1,3}(F_1(x_1), F_3(x_3))f_3(x_3)$.

Each conditional density function in equation (29) can be decomposed into a product of the corresponding pair-copulas using a general formula

$$f(x_i|\mathbf{v}) = c_{x_i, x_j|\mathbf{v}_{-j}}(F(x_i|\mathbf{v}_{-j}), F(x_j|\mathbf{v}_{-j})|\mathbf{v}_{-j})f(x_i|\mathbf{v}_{-j}) \quad (36)$$

for $i, j = 1, \dots, n$, and \mathbf{v} represents any set of x_1, \dots, x_n where x_j is in the set, but there is no x_i . Then, \mathbf{v}_{-j} is an n -dimensional vector with the exclusion of the j -th component [1].

Next, the construction of three-dimensional pair-copulas will be illustrated. For three random variables X_1, X_2 , and X_3 , the joint density function can be decomposed as follows

$$f(x_1, x_2, x_3) = f(x_3|x_1, x_2)f(x_2|x_1)f_1(x_1) \quad (37)$$

The conditional density function $f(x_2|x_1)$ can be expressed as in equation (31). The conditional density function $f(x_3|x_1, x_2)$ can also be expressed as in equation (34). Then, by substituting equations (31) and (35) into equation (37), one obtains one form of the density function of the three-dimensional pair-copula decomposition

$$f(x_1, x_2, x_3) = c_{2,3|1}(F(x_2|x_1), F(x_3|x_1)|x_1)c_{1,3}(F_1(x_1), F_3(x_3)) \quad (38)$$

$$c_{1,2}(F_1(x_1), F_2(x_2))f_3(x_3)f_2(x_2)f_1(x_1)$$

This PCC decomposition is not unique because if different conditional variables are used in equations (31) and (35), it would result in different pair-copula constructions. Analogously, other decompositions of $f(x_1, x_2, x_3)$ can be as follows

$$f(x_1, x_2, x_3) = c_{1,3|2}(F(x_1|x_2), F(x_3|x_2)|x_2)c_{2,3}(F_2(x_2), F_3(x_3)) \quad (39)$$



$$c_{1,2}(F_1(x_1), F_2(x_2))f_3(x_3)f_2(x_2)f_1(x_1)$$

and

$$f(x_1, x_2, x_3) = c_{1,2|3}(F(x_1|x_3), F(x_2|x_3)|x_3)c_{1,3}(F_1(x_1), F_3(x_3))c_{2,3}(F_2(x_2), F_3(x_3))f_3(x_3)f_2(x_2)f_1(x_1) \tag{40}$$

Suppose there is a conditional copula $c_{i,j|D}(\cdot, \cdot | \mathbf{x}_D)$. To simplify the estimation process, a simplifying assumption is used, where the conditioning variable x_D is ignored. Therefore, it holds that $c_{i,j|D}(\cdot, \cdot | \mathbf{x}_D) = c_{i,j|D}(\cdot, \cdot)$. Subsequently, the simplifying assumption applies to vine copulas [8].

5.2. Regular Vine Copula

In high dimensions, there are many possible constructions of pair-copulas. In 2001 and 2002, Bedford and Cooke [3][4] developed a graphical structure called a regular vine tree sequence to characterize and organize all factorizations. A set of trees $\mathcal{V} = (T_1, \dots, T_{d-1})$ is a regular vine tree sequence on d elements if:

- Each tree $T_j = (N_j, E_j)$ is connected, where $j = 1, \dots, d - 1$.
- T_1 is a tree with node set $N_1 = \{1, \dots, d\}$ and edge set E_1 .
- For $j \geq 2$, T_j is a tree with node set $N_j = E_{j-1}$ and edge set E_j
- For $j = 2, \dots, d - 1$, for $a, b \in N_{j-1}$ and $\{a, b\} \in E_j$, it must hold that $|a \cap b| = 1$ (proximity condition).

As an example, here is a regular vine tree sequence for the construction of a three-dimensional pair-copula, as discussed earlier.

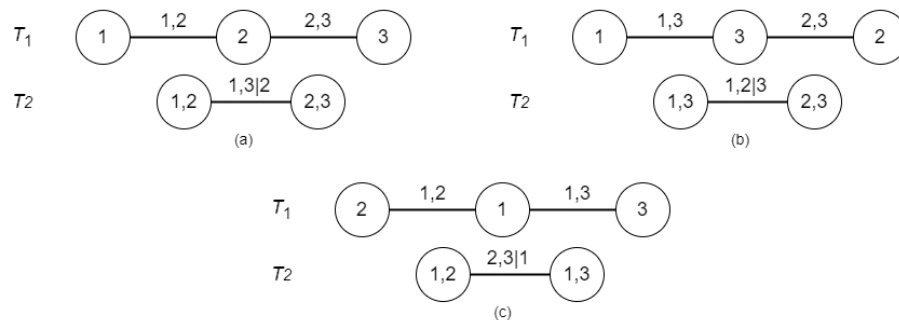


Figure 1. Vine Copula Structure in Three Dimension

Regular vines have two subclasses: canonical vines and drawable vines. A regular vine tree sequence $\mathcal{V} = (T_1, \dots, T_{d-1})$ is called a C-vine tree sequence if, for each tree T_i , there exists a node $n \in N_i$ such that $|\{e \in E_i | n \in e\}| = d - i$. This node is referred to as the root node of the tree T_i . A regular vine tree sequence $V = (T_1, \dots, T_{d-1})$ is called a D-vine tree sequence if, for every node $n \in N_i, |\{e \in E_i | n \in e\}| \leq 2$. According to Czado [8], the density function of a regular vine copula is as follows

$$f(x_1, \dots, x_d) = \left[\prod_{k=1}^d f_k(x_k) \right] \left[\prod_{k=1}^{d-1} \prod_{e \in E_k} c_{i,j|D} (F_{i|D}(x_i | \mathbf{x}_D), F_{j|D}(x_j | \mathbf{x}_D)) \right] \tag{41}$$

For canonical vine copulas, the density function is as follows

$$f(x_1, \dots, x_d) = \left[\prod_{k=1}^d f_k(x_k) \right] \left[\prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{j,j+i|1, \dots, j-1} \right] \tag{42}$$

And for drawable vine copulas, the density function is as follows



$$f(x_1, \dots, x_d) = \left[\prod_{k=1}^d f_k(x_k) \right] \left[\prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{i,i+j|i+1,\dots,i+j-1} \right] \quad (43)$$

6. Pseudo-Maximum Likelihood

Pseudo-Maximum Likelihood Method is used to estimate parameters of copula models when the marginal distribution functions of the data are unknown. This method is similar to Maximum Likelihood Estimation (MLE) and allows the estimation of copula parameters based on the likelihood function of its marginal distribution. The Pseudo-Maximum Likelihood method was introduced by Genest in 1995 [8].

Let there be a random two-dimensional sample of size n , denoted as (X_1, X_2, \dots, X_n) , which are mutually independent and identically distributed, where $X_j = (X_{1j}, X_{2j})$ for $j = 1, \dots, n$. Then, the n data vectors are transformed into pseudo-observations by calculating the empirical distribution function of each marginal, resulting in $(\widehat{F}_{1n}(x_{1j}), \widehat{F}_{2n}(x_{2j}))$.

In this method, copula parameters are estimated by maximizing the pseudo-likelihood function. According to Sklar's Theorem, for the case of two variables, it can be expressed as $H(x_1, x_2) = C(F(x_1), F(x_2))$. Let $L(\theta; x)$ be the pseudo-likelihood function, then the parameters will be estimated by maximizing the pseudo-likelihood function:

$$\hat{\theta} = \arg \max \sum_{j=1}^n \ln c(\widehat{F}_{1n}(x_{1j}), \widehat{F}_{2n}(x_{2j}); \theta) \quad (44)$$

7. Sequential Estimation

Within this section, the discussion will revolve around the sequential estimation of copula parameters within a vine tree sequence. This estimation involves a sequential order of parameter estimation, starting from the first tree, then the second tree, and so on until the last tree.

Let there be copula parameters denoted as θ_e , where edge $e = (a_e, b_e | D_e)$ is part of a regular vine tree sequence in tree T_i . The copula parameters for this edge are denoted as $\theta(T_i)$, and their estimates are represented as $\hat{\theta}(T_i)$. Assume that all pair-copula parameters in trees from T_1 to T_{i-1} have already been estimated. The set of estimated parameters is denoted as $\hat{\theta}(T_{1,\dots,i-1})$.

The sequential estimation of θ_e , for edge $e = (a_e, b_e | D_e)$ in tree T_i is based on pseudo-observations. These pseudo-observations are defined as follows:

$$u_{k,a_e|D_e,\hat{\theta}(T_{1,\dots,i-1})} = C_{a_e|D_e}(u_{k,a_e} | \mathbf{u}_{k,D_e}, \hat{\theta}(T_{1,\dots,i-1})) \quad (45)$$

$$u_{k,b_e|D_e,\hat{\theta}(T_{1,\dots,i-1})} = C_{b_e|D_e}(u_{k,b_e} | \mathbf{u}_{k,D_e}, \hat{\theta}(T_{1,\dots,i-1})) \quad (46)$$

where $k = 1, \dots, n$. These pseudo-observations are used to estimate θ_e . Utilizing these pseudo-observations θ_e , is estimated by seeking values that maximize the product of the copula function $c_{a_e,b_e|D_e}$ based on these pseudo-observations [8]. In essence, the objective is to identify the values of θ_e that maximize the following expression

$$\prod_{k=1}^n c_{a_e,b_e|D_e}(u_{k,a_e|D_e,\hat{\theta}(T_{1,\dots,i-1})}, u_{k,b_e|D_e,\hat{\theta}(T_{1,\dots,i-1})}; \theta_e) \quad (47)$$

8. Akaike Information Criterion

Akaike Information Criterion (AIC) is a method used to evaluate how well a statistical model fits a given dataset. AIC is commonly used for model selection and comparison by estimating the quality of each model relative to other models. AIC was developed by the Japanese statistician Hirotugu Akaike. The AIC value is defined as:

$$AIC = -2 \ln L(\hat{\theta}) + 2k \quad (48)$$



Where θ is the set of model parameters, $L(\hat{\theta})$ is the maximum likelihood estimation of θ , and k is the number of estimated parameters. The best model is the one with the lowest AIC value [13].

9. Empirical Study

This section explains the application of the vine copula model by describing the utilized dataset. The process begins by first transforming the data and then selecting copula families and estimating parameters for the edge of the first tree in each model. Pseudo-observations are then formed to estimate parameters for the edge of the second tree, followed by the selection of copula families and parameter estimation for the second tree. After that, model selection is performed.

9.1. Data Description

The application of the vine copula model using uranium exploration data obtained from a study conducted by Cook and Johnson in 1986. The dataset consists of information about the logarithm concentrations of chemical elements in 655 water samples collected around Grand Junction, Colorado. Caesium, Scandium, and Titanium elements will be analyzed in this study using the vine copula model. The dataset involves the concentrations of multiple chemical elements and environmental data often exhibit complex dependencies due to geological, hydrological, or ecological factors. Vine copulas are well-suited for modeling the joint distribution of multiple variables and useful when there is a need to capture complex dependence structures between variables.

9.2. Data Transformation to Pseudo-Observations

Data will be transformed into pseudo-observations. The empirical probability integral transform is employed for this purpose. This transformation ranks each data point and utilizes the empirical distribution function on the transformed rank data, mathematically expressed as $u_{ij} = \frac{r_{ij}}{n+1} = \frac{r_{ij}}{656}$ for each element in each vector $x_i = (x_{i1}, \dots, x_{id})$, where u_{ij} is the pseudo-observation, r_{ij} is the data rank, and n is the data count. This transformation results in uniformly distributed values between 0 and 1.

9.3. Fitting Model and Parameter Estimation

Within this section, model fitting and the estimation of parameters will be executed for the dependency structure contained within the vine copula. In this case, there are three variables that need to be modeled, resulting in three possible vine copula structures, as shown in Figure 1. For simplicity, the vine copula structures in Figure 1 (a), (b), and (c) will be referred to as models 1, 2, and 3, respectively.

The search for the most appropriate bivariate copula and the determination of the respective parameters for each pair copula within all potential dependency structures will be conducted. Figure 2 provides a visual overview of the relationship between the variables to be modeled.

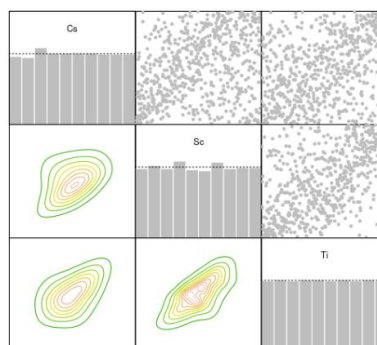


Figure 2. Contour Plot, Histogram, and Scatter Plot of Pseudo-Observation



The selection of the best copula family for $c_{1,2}$, $c_{1,3}$, and $c_{2,3}$ will be done using the Akaike Information Criterion (AIC). The results are as follows:

Table 1. AIC for Each Copula Family for $c_{1,2}$, $c_{1,3}$, and $c_{2,3}$

Copula family	AIC $c_{1,2}$	AIC $c_{1,3}$	AIC $c_{2,3}$
Gaussian	-77	-143	-295
Student's t	-114	-170	-315
Clayton	-100	-71	-238
Gumbel	-65	-186	-286
Frank	-74	-140	-298

From the table above, for $c_{1,2}$, the smallest AIC value is shown by the Student's t copula family. Then, for $c_{1,3}$, the smallest AIC value is indicated by the Gumbel copula family, and for $c_{2,3}$, the smallest AIC value is shown by the Student's t copula family. After selecting the copula families, the next step is to estimate the parameters for each copula. Parameter estimation will be performed using the sequential method and pseudo-maximum likelihood method. Parameter estimation will be carried out for the first set of copulas in each model's structure, followed by parameter estimation for the second set of copulas by creating pseudo-observations using the estimated parameters from the first set. Using the pseudo-maximum likelihood method, the parameter estimates are as follows:

$$\begin{aligned} \hat{\theta}_{1,2} &= (0.34, 3.48) \\ \hat{\theta}_{1,3} &= 1.5 \\ \hat{\theta}_{2,3} &= (0.62, 5.93) \end{aligned}$$

In the following steps, the copula family for the second set of copulas will be determined. In Model 1, pseudo-observations $u_{1|2} = c(u_1|u_2; \hat{\theta}_{1,2})$ and $u_{3|2} = c(u_3|u_2; \hat{\theta}_{2,3})$ are generated to estimate the parameters of $c_{1,3|2} = (u_{1|2}, u_{3|2}; \theta_{1,3|2})$. Similar steps are applied for Models 2 and 3. Visual overview of the contour plots for $c_{1,3|2}$, $c_{1,2|3}$ and $c_{2,3|1}$ is presented in Figure 3.

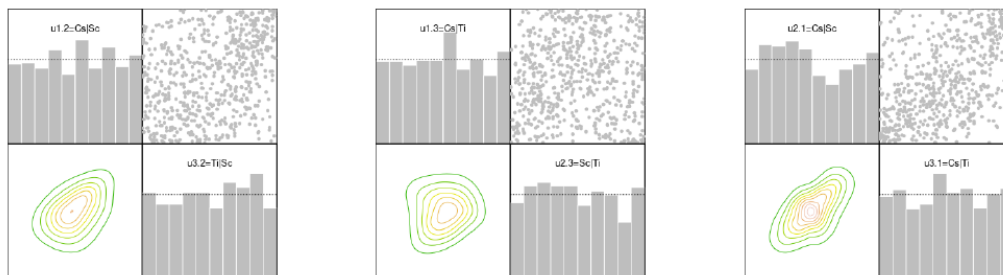


Figure 3. Contour Plots for Conditional Copulas (Left: $u_{1|2}$ and $u_{3|2}$, Middle: $u_{1|3}$ and $u_{2|3}$, Right: $u_{2|1}$ and $u_{3|1}$)

To objectively select the copula family, the Akaike Information Criterion (AIC) is used, and the results are shown in Table 2:

Table 2. AIC for Each Copula Family for $c_{1,3|2}$, $c_{1,2|3}$ and $c_{2,3|1}$

Copula Family	AIC $c_{1,3 2}$	AIC $c_{1,2 3}$	AIC $c_{2,3 1}$
Gaussian	-81	-1.46	-212
Student's t	-83	-14.43	-230
Clayton	-26	-15.64	-70
Gumbel	-91	-1.75	-234
Frank	-69	-1.42	-210

From the table above, the copula family with the smallest AIC value was selected. For $c_{1,3|2}$, the Gumbel copula family was chosen. Then, for $c_{1,2|3}$ and $c_{2,3|1}$, the Student's t copula family was selected.



After determining the copula families, the next step is to estimate the parameters of each copula using the pseudo-maximum likelihood method. Subsequently, the tail dependency parameters of the copulas with Gumbel and Student's t copula families will be calculated. All the obtained parameters for each model are presented in the following table.

Table 3. Model Fitting and Parameter Estimation Results

Pair Copula	Copula Family	Parameter 1	Parameter 2	Upper Tail Dependence	Lower Tail Dependence	AIC
Model 1						
(Cs, Sc)	Student's t	0.34	3.48	0.21	0.21	-532
(Sc, Ti)	Student's t	0.62	5.93	0.25	0.25	
(Cs, Ti Sc)	Gumbel	1.30	0.00	0.29	-	
Model 2						
(Cs, Ti)	Gumbel	1.46	0.00	0.39	-	-515
(Sc, Ti)	Student's t	0.62	5.93	0.25	0.25	
(Cs, Sc Ti)	Student's t	0.08	6.85	0.03	0.03	
Model 3						
(Cs, Sc)	Student's t	0.34	3.48	0.21	0.21	-529
(Cs, Ti)	Gumbel	1.46	0.00	0.39	-	
(Sc, Ti Cs)	Student's t	0.55	6.06	0.20	0.20	

The table above summarizes the results of model fitting and parameter estimation for the three models.

As for the model structure, Model 1 exhibited the lowest AIC value. Therefore, Model 1 was selected as the best model. The vine copula model for the dependencies between Caesium, Scandium, and Titanium is as follows:

$$\begin{aligned}
 f(x_1, x_2, x_3) = & \left(\frac{\left((-\ln F(x_1|x_2))(-\ln F(x_3|x_2)) \right)^{0,3}}{F(x_1|x_2)F(x_3|x_2)} \exp \left(- \left((-\ln F(x_1|x_2))^{1,3} \right. \right. \right. \\
 & \left. \left. \left. + (-\ln F(x_3|x_2))^{1,3} \right) \right)^{0,769231} \left((0,3) \left((-\ln F(x_1|x_2))^{1,3} \right. \right. \right. \\
 & \left. \left. \left. + (-\ln F(x_3|x_2))^{1,3} \right)^{-1,230769} \right. \right. \\
 & \left. \left. \left. + \left((-\ln F(x_1|x_2))^{1,3} + (-\ln F(x_3|x_2))^{1,3} \right)^{-0,461538} \right) \right) \\
 & \left(\frac{0,202848[1 + 0,273934(x_2 - 1,24x_2x_3 + x_3^2)]^{-3,965}}{0,146342 \left[\left(1 + \frac{x_2^2}{5,93} \right) \left(1 + \frac{x_3^2}{5,93} \right) \right]^{-3,465}} \right) \\
 & \left(\frac{0,169237[1 + 0,324917(x_1 - 0,68x_1x_2 + x_2^2)]^{-2,74}}{0,138105 \left[\left(1 + \frac{x_1^2}{3,48} \right) \left(1 + \frac{x_2^2}{3,48} \right) \right]^{-2,24}} \right) f_1(x_1)f_2(x_2)f_3(x_3)
 \end{aligned}$$

Model 1 suggests that there is a dependence relationship between Caesium and Titanium with Scandium as a conditional variable. The dependence between Caesium and Scandium is modeled using the Student's t copula, indicating that there is dependence when both values are very large or very small. The dependence parameter between these two elements is relatively low at 0.34, indicating that their dependence is not very strong. The Student's t copula used in this model has a degree of freedom of



3.48. The upper and lower tail dependence in this model is not too strong, with upper and lower tail dependence parameters of 0.21.

Additionally, the dependence relationship between Scandium and Titanium is also modeled using the Student's *t* copula. This suggests that there is dependence between these two elements when their values are very large or very small. The dependence between these two elements is quite significant, with a dependence parameter of 0.62. The degree of freedom for this model is 5.93, which is higher than the Student's *t* copula for Caesium and Scandium. This results in relatively weak tail dependence, despite having a relatively large dependence parameter of 0.25.

Furthermore, the dependence relationship between Caesium and Titanium in the presence of Scandium is modeled using the Gumbel copula with a parameter of 1.30. This indicates that the level of dependence between them tends to be low because the parameter approaches 1. In this model, there is dependence when the values of both elements are very large, with an upper tail dependence parameter of 0.29. Although this dependence is relatively weak, it is stronger than the tail dependence of other pair-copulas.

Among these three dependencies, Scandium and Titanium exhibit the highest dependence. From a chemical perspective, these two elements are not directly related and do not interact with each other in water. Generally, regions with high scandium concentrations tend to also have high titanium concentrations due to both elements being commonly found together in the same ore deposits [7]. The concentration of caesium, titanium, and scandium in water can have diverse and significant effects on the environment. Scandium, for instance, has the potential to gradually accumulate in soils and water, posing risks to both human and animal health, particularly when released into the environment by various industries. Titanium, on the other hand, exhibits remarkable resistance to corrosion by seawater, making it a valuable material for an array of ocean-related applications, such as propeller shafts, rigging, and desalination plants. Caesium, in contrast, is known for its explosive reaction with water, which can lead to ignition and violent explosions, presenting clear safety hazards [22].

10. Conclusion

The construction of pair-copulas can be carried out through the decomposition of conditional probability density functions and substituting copula density functions into the resulting decomposition. Thus, a multivariate density function can be formed, consisting of bivariate copula density functions. Subsequently, this probability density function can be organized by a graphical structure called a "vine." This graphical structure consists of a set of trees.

In this research, parameters were estimated using the pseudo-maximum likelihood method and sequential estimation. The pseudo-maximum likelihood method is similar to the maximum likelihood method, but the marginal distribution is unknown. Therefore, data is transformed into pseudo-observations, and then parameters that maximize its likelihood function are sought. Estimation in the vine copula is done sequentially, starting from the first tree, using parameters from the first tree to create pseudo-observations used to estimate the next tree, and so on until the last tree is reached.

The application of vine copulas in this research used empirical data with unknown marginal distributions. Therefore, data transformation into pseudo-observations was performed. For each possible vine structure, copula families were selected using the Akaike information criteria. Then, parameter estimation was carried out using sequential estimation and pseudo-maximum likelihood for all bivariate copulas. After that, the best vine structure that can model data dependency based on the Akaike information criteria was selected. The results showed that Caesium and Titanium have a dependency relationship on Scandium. The dependence between Scandium and Titanium is the strongest compared to other variable pairs.

These findings emphasize the need for comprehensive monitoring and regulation of these chemical elements in water sources to safeguard both environmental and human well-being. By recognizing the statistical associations between these elements, authorities can pinpoint potential sources of contamination or natural geological factors impacting water composition. Moreover, the correlation between high scandium and titanium concentrations, attributed to their common presence in the same



ore deposits, underscores the significance of responsible resource management. This insight can inform resource extraction and environmental protection policies in areas rich in these minerals. Consequently, policymakers and environmental agencies can make informed decisions using data analysis, particularly in cases where regulations or interventions are required to safeguard water quality and mitigate the consequences of mining and industrial activities.

For further research, in higher dimensions, manually searching for an appropriate vine copula structure and family can become highly challenging and time-consuming. Therefore, it may be worth considering the use of Dißmann's algorithm for a more efficient model fitting of vine copulas.

References

- [1] Aas K, Czado C, Frigessi A, Bakken, H 2009 Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics* **44(2)** pp 182-198
- [2] AghaKouchak A, Sellars S, Sorooshian, 2012 Methods of Tail Dependence Estimation *Extremes in a Changing Climate* pp 163–179
- [3] Bedford T, Cooke R M 2001 Monte Carlo simulation of vine dependent random variables for applications in uncertainty analysis *In Proceedings of ESREL2001 (Italy: Turin)*
- [4] Bedford T, Cooke R M 2002 Vines: A new graphical model for dependent random variables *Annals of Statistics* **30(4)** pp 1031–1068
- [5] Brechmann E C, Czado C 2013 Risk management with high-dimensional vine copulas: An analysis of the Euro Stoxx 50 *Statistics & Risk Modeling* **30(4)**
- [6] Cook R, Johnson M 1986 Generalized Burr-Pareto-logistic distributions with applications to a uranium exploration data set *Technometrics* **28(2)** pp 123–131
- [7] Cui T, Nie A 2018 Geological Features of Scandium Deposits in Southwestern Guizhou Province *IOP Conference Series: Earth and Environmental Science* **170**
- [8] Czado C 2019 *Analyzing Dependent Data with Vine Copulas : A Practical Guide With R* (Springer International Publishing)
- [9] Dißmann J F 2010 Statistical inference for regular vines and application (Technische Universität München)
- [10] Dorje O T 2018 *INTERCONNECTED : embracing life in our global society* (Wisdom Publications)
- [11] Embrechts P, Lindskog F, Mcneil A 2003 Modelling Dependence with Copulas and Applications to Risk Management *Handbook of Heavy Tailed Distributions in Finance* pp 329–384
- [12] Erhardt V, Czado C 2012 Modeling dependent yearly claim totals including zero claims in private health insurance *Scandinavian Actuarial Journal* **2012(2)** pp 106–129.
- [13] Fabozzi F J, Focardi S M, Rachev S T, Arshanapalli B G 2014 *The Basics of Financial Econometrics Tools, Concepts, and Asset Management Applications* (Hoboken, Nj, Usa John Wiley & Sons, Inc.)
- [14] Hadiputra F F 2020 Pelabelan total super simpul antiajaib lokal pada graf (Universitas Indonesia)
- [15] Hoeffding W 1940 Scale-invariant correlation theory. *N. I. Fisher and P. K. Sen (eds.): The Collected Works of Wassily Hoeffding* (New York: Springer-Verlag) pp 57–107
- [16] Hoeffding W 1941 Scale-invariant correlation measures for discontinuous distributions *N. I. Fisher and P. K. Sen (eds.): The Collected Works of Wassily Hoeffding* (New York: Springer-Verlag) pp 109–133
- [17] Hogg R V, Mckean J W, Craig A T 2004 *Introduction to mathematical statistics* (Pearson Education)
- [18] Hohndorf L, Czado C, Bian H, Kneer J, Holzapfel F 2017 Statistical modeling of dependence structures of operational flight data measurements not fulfilling the iid condition. *In AIAA Atmospheric Flight Mechanics Conference* p. 3395
- [19] Horvath G, Kovacs E, Molontay R, Novaczki, S 2020 Copula-Based Anomaly Scoring and Localization for Large-Scale, High-Dimensional Continuous Data *ACM Transactions on Intelligent Systems and Technology* **11(3)** pp 1–26



- [20] Jeong H, Dey D 2019 *Application of vine copula for multi-line insurance reserving*
- [21] Joe, H 1996 Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters In L. Ruschendorf, B. Schweizer, & M. D. Taylor (Eds.), *Distributions with Fixed Marginals and Related Topics*.
- [22] Lennetch, "Elements," *Lenntech.com.pt*, 2017.
<https://www.lenntech.com.pt/periodico/elements.htm> (accessed Oct. 28, 2023).
- [23] Markowitz H 1952 Portfolio Selection *The Journal of Finance* **7(1)** pp 77–91.
- [24] Nelsen R B 2007 *An Introduction to Copulas* (Springer Science & Business Media)
- [25] Nikoloulopoulos A K 2017 A vine copula mixed effect model for trivariate metaanalysis of diagnostic test accuracy studies accounting for disease prevalence *Statistical Methods in Medical Research* **26(5)** pp 2270–2286
- [26] Omari C, Mwita P, Waititu A 2019 *Conditional Dependence Modelling with Regular Vine Copulas* **8** pp 97-133
- [27] Rosen K H 2017 *Discrete Mathematics and Its Applications* (McGraw-Hill)
- [28] Trivedi P K, Zimmer D M 2009 Pitfalls in Modelling Dependence Structures: Explorations with Copulas* *The Methodology and Practice of Econometrics* pp 149–172
- [29] Venter G 2001 Tails of copulas *Proceedings ASTIN Washington* pp 68-113
- [30] Walpole R E, Myers R H, Myers S L, Ye K 2012 *Probability & statistics for engineers & scientists* (Prentice Hall)