On the use of copulas in geotechnical engineering: A tutorial and state-of-the-art-review

Juan J. Sepúlveda-García[1*] and Diego A. Alvarez[2]

[1]Departamento de Ingeniería Civil y Agrícola, Universidad Nacional de Colombia, sede Bogotá, Cr. 45 # 26-85, Bogotá D.C., 111321, Colombia.

[2]Departamento de Ingeniería Civil, Universidad Nacional de Colombia, sede Manizales, Cr. 27 # 64-60, Manizales, 170004, Colombia.

Abstract

Copulas are functions that couple marginal distribution in order to generate joint probability distribution functions; in this way, they model the dependence among random variables. This paper presents a tutorial and an state-of-the-art review that exposes many applications of copula theory in the field of geotechnical engineering, discussing its advantages, drawbacks, limitations, among other aspects, based on the existing literature. The results are conclusive about the importance and necessity of adoption of copula theory in the field of geotechnical engineering.

Keywords: copulas, reliability analysis, dependence, joint probability distribution functions, correlation

This in an electronic reprint of the original article published by Springer in Archives of Computational Methods in Engineering (2022), DOI: https://doi.org/10.1007/s11831-022-09771-2 (please cite accordingly). This reprint differs from the original in pagination and typographic details.

1 Introduction

Uncertainty, in geotechnical engineering, is commonly handled, in an implicit way, by employing factors of safety and partial factors of safety. However, more advanced methods employ a probabilistic description of the system that represent all involved quantities such as shear strength, unit weight or peak ground acceleration, among others, as random variables. Those (marginal) random variables X_i might be grouped into a random vector $X = [X_1, X_2, \ldots, X_d]$ which follows a joint cumulative distribution function (CDF) F_X or, alternatively, a joint probability distribution function (PDF) f_X.

Unfortunately, it is usual that the only probabilistic description of the set of involved random variables X_i is their marginal PDFs and a correlation matrix; hence, a joint PDF cannot be determined uniquely [see, e.g., 1, 2]. This type of representation leads to assumptions, that are not only weakly supported but also generate errors in the results, that discourage the acceptance of probabilistic methods in practice. Some of the most usual assumptions that are chosen in order to simplify calculations are: the representation of variables as Gaussian without any evidence, the supposition that Pearson's rho coefficient can accurately model the correlation between variables, or the assumption that variables are independent even if they were obtained from the same sample and/or the same laboratory test, like in the case of the shear strength parameters, c and ϕ. Attempts to overcome this disadvantage have been developed, being one of the most notable the Nataf transformation [3], which will be discussed in section 3.1.

Instead of using unfounded assumptions, copula theory, introduced by the american mathematician Abe Sklar in 1959 [4], offers a comprehensive framework on dependence and construction of multivariate probability models and it is currently consolidated as the dependence theory par excellence. A copula is a function that joins or couples several marginal CDFs with the aim of constructing a multivariate CDFs. Thus, a copula models the dependence information in a joint CDF.

Copula theory has been widely applied in areas like actuary and finance [see, e.g., 5, 6], hydrology [see, e.g., 7, 8, 9], structural engineering [see, e.g., 10], and more recently, in geotechnics. For instance, copulas have been used in geotechnical engineering for modeling dependence among shear strength parameters, to estimate settlement of foundations, or representing different soil properties.

Nonetheless, although copula theory has been used in different contexts and problems of geotechnical engineering, these works

*Corresponding author
Email addresses: jsepulveda@unal.edu.co (Juan J. Sepúlveda-García), daalvarez@unal.edu.co (Diego A. Alvarez)

Juan J. Sepúlveda-García and Diego A. Alvarez have contributed equally to this work.

have been rather sparse and independent. Therefore, there is currently no methodology or framework that guides students, researchers, or practicing engineers in the area of geotechnics on the application of copulas in their particular problems. Thus, this article aims to fill this gap, in the state-of-the-art, by introducing copula theory while highlighting its potential, scope, limitations and use in geotechnical engineering.

In this paper questions like: do we necessarily have to model dependence in geotechnics?, which applications of copulas in geotechnics exist?, do copula models perform better than the traditionally Nataf transformation?, how do the gaussianity or independence assumptions perform concerning copulas?, is there “the most suitable” copula for geotechnical models?, among others, will be answered.

The plan of this document is the following: section 2 introduces copula theory; section 3 exposes how dependence has been commonly addressed in geotechnics, and how traditional approaches to deal with dependence have been applied; section 4 explains the basis for the construction of copulas with an emphasis in geotechnical models; in section 5, an state-of-the-art on the applicability of copula theory in geotechnical engineering is presented; section 6 deals with the topic of uncertainties in the copula construction and selection; in section 7, the applicability of copulas in reliability analysis of geotechnical models is studied and a practical example is presented; the document ends in section 8 with some conclusions and open questions in the state-of-the-art.

All data and codes used in this manuscript are publicly available on GitHub at https://github.com/juanjosesg/geo-copulas.git.

2 Introduction to copula theory

This section introduces copula theory, mainly focused on those practical concepts that are needed in order to apply copulas in the context of geotechnical engineering, after Nelsen [11], Embrechts et al. [12, 13] and McNeil et al. [5].

2.1 Preliminary concepts

This section presents some preliminary concepts that will help us later, when the copulas are formally defined.

2.1.1 Increasing functions

A function $f : \mathbb{R} \mapsto \mathbb{R}$ is called *increasing* if for all $x \leq y$ the relationship $f(x) \leq f(y)$ is satisfied. Sometimes these functions are also referred as *non-decreasing*. Furthermore, the function f is said to be *strictly increasing* if for all $x < y$ the relationship $f(x) < f(y)$ holds. Analogously, this function is also referred as *strictly non-decreasing* function.

On the other hand, if the increasing (or strictly increasing) function f is differentiable on its domain, then the function f is also said to be *monotonic*.

2.1.2 Quasi-inverses

The concept of *quasi-inverse* is a generalization of the concept of the inverse of a function, in which the inverse is defined even when the function has jumps or discontinuities. Let us consider a non-decreasing function $F : \mathbb{R} \to R$ with $R \subseteq \mathbb{R}$, and the extended real line $\bar{\mathbb{R}} = [-\infty, \infty]$. Then, a quasi-inverse of F is any function $F^{(-1)} : R \to \bar{\mathbb{R}}$, such that:

- If $t \in R$, then $F^{(-1)}(t)$ is any number x in $\bar{\mathbb{R}}$ such that $F(x) = t$.
- If $t \notin R$, then $F^{(-1)}(t) = \inf\{x \in \bar{\mathbb{R}} | F(x) \geq t\}$.

Note that the concept of quasi-inverse and inverse of a function coincide when F is a strictly monotonically increasing function; in this case, the inverse is denoted as F^{-1}.

2.1.3 V_H-volume

Let $S_i = [a_i, b_i]$ with $i = 1, 2, \ldots, d$, be non-empty subsets of $\bar{\mathbb{R}}$, and let H be a mapping $H : S \to \mathbb{R}$. Furthermore, let $B = \times_{i=1}^{d} S_i$ be a d-dimensional box completely contained on S, that is, $B \subseteq S$. The H-volume of B, $V_H(B)$, is defined as the *dth-order difference* of H on B, as follows:

$$V_H(B) = \Delta_{a_d}^{b_d} \Delta_{a_{d-1}}^{b_{d-1}} \ldots \Delta_{a_1}^{b_1} H$$

where the first order differences of H are defined, for $k = 1, \ldots, d$, as:

$$\Delta_{a_k}^{b_k} H = H(t_1, \ldots, t_{k-1}, b_k, t_{k+1}, \ldots, t_d) \\ - H(t_1, \ldots, t_{k-1}, a_k, t_{k+1}, \ldots, t_d).$$

This expression can be rewritten more explicitly as:

$$V_H(B) = \sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{i_1+\cdots+i_d} H\left(x_{1 i_d}, \ldots, x_{d i_d}\right)$$

where $x_{j1} = a_j$ and $x_{j2} = b_j$ for all $j = 1, \ldots, d$ [see, e.g., 13].

For instance, in the bidimensional case, $d = 2$ and $B = [a_1, b_1] \times [a_2, b_2]$, and hence the H-volume of the rectangle B is the second order difference of H on B:

$$V_H(B) = H(a_2, b_2) - H(a_1, b_2) - H(a_2, b_1) + H(a_1, b_1).$$

We will say that the d-dimensional function H is *d-increasing* if $V_H(B) \geq 0$ for all d-boxes whose vertices fall in the domain of H.

Note that when H is a multivariate CDF, the V_H-volume of a d-dimensional box B is basically, the probability measure associated to B.

2.1.4 Multivariate cumulative distribution functions

A *d-dimensional (multivariate or joint) cumulative distribution function* is defined as a function $H : \bar{\mathbb{R}} \to [0, 1]$ that fulfills the following properties:

- H is d-increasing.
- $H(t) = 0$ if $t = [t_1, \ldots, t_k, \ldots, t_d] \in \bar{\mathbb{R}}$ has any component $t_k = -\infty$, for $k = 1, \ldots, d$.
- $H(\infty, \ldots, \infty) = 1$

2.1.5 k-margins

Let us consider the joint cumulative distribution function $H : S_1 \times S_2 \times \cdots \times S_d \to [0,1]$. Then, the function

$$H_j(x) = H\big(\sup S_1, \ldots, \sup S_{j-1}, x, \sup S_{j+1}, \ldots, \sup S_d\big)$$

is a univariate margin of H, or simply a margin of H. Note that higher dimensional margins (k-margins, $k \geq 2$) can be defined by fixing less positions in the previous equation.

2.2 Copulas

Copulas are functions that couple several CDFs through a dependence structure in order to represent a multivariate CDF. In this sense, one can characterize and define a marginal CDF for each random variable, and then model their dependence with copulas.

Mathematically, a copula $\mathtt{C}$ is a d-dimensional CDF whose domain is $[0,1]^d$ and whose range is $[0,1]$, that is, $\mathtt{C} : [0,1]^d \to [0,1]$. In this regard, every marginal of a d-dimensional copula is uniform in the interval $[0,1]$. Furthermore, any copula function $\mathtt{C}$ has the following properties:

1. $\mathtt{C}$ is *grounded*, that is, for all $u \in [0,1]^d$, then $\mathtt{C}(u) = 0$ if there exists any $u_i = 0$, for $i = 1, \ldots, d$.
2. $\mathtt{C}$ has *margins*, i.e., for all $u \in [0,1]^d$, then $\mathtt{C}(u) = u_i$ if all elements of u are 1 except u_i.
3. $\mathtt{C}$ is *d-increasing*, that is, for all $[a_1, \ldots, a_i, \ldots, a_d], [b_1, \ldots, b_i, \ldots, b_d] \in [0,1]^d$, and taking into account that $a_i \leq b_i$ for $i = 1, \ldots, d$, then $V_{\mathtt{C}}\big(\times_{i=1}^{d} [a_i, b_i]\big) \geq 0$.

For instance, if the copula is bidimensional ($d = 2$) then $B = [a_1, b_1] \times [a_2, b_2]$ and hence the condition to be 2-increasing becomes:

$$\mathtt{C}(a_2, b_2) - \mathtt{C}(a_1, b_2) - \mathtt{C}(a_2, b_1) + \mathtt{C}(a_1, b_1) \geq 0$$

By virtue of the Lebesgue's decomposition theorem, every d-dimensional copula can be decomposed as [see, e.g., 11]:

$$\mathtt{C}(u_1, \ldots u_d) = A_{\mathtt{C}}(u_1, \ldots, u_d) + S_{\mathtt{C}}(u_1, \ldots, u_d)$$

where A_C represents the *absolutely continuous component* of the copula:

$$A_{\mathtt{C}}(u_1, \ldots, u_d) = \int_0^{u_1} \cdots \int_0^{u_d} \frac{\partial^d}{\partial u_1, \ldots, \partial u_d} \mathtt{C}(u_1, \ldots, u_d)\, \mathrm{d}s_1 \cdots \mathrm{d}s_d$$

and S_C is its *singular component*:

$$S_{\mathtt{C}}(u_1, \ldots, u_d) = \mathtt{C}(u_1, \ldots u_d) - A_{\mathtt{C}}(u_1, \ldots, u_d).$$

When $S_{\mathtt{C}} = 0$ on $[0,1]^d$, then the copula $\mathtt{C}$ is said to be *absolutely continuous* and its joint PDF, also known as *copula density*, $\mathtt{c}$ exists and it is defined as:

$$\mathtt{c}(u_1, \ldots u_d) = \frac{\partial^d}{\partial u_1, \ldots, \partial u_d} \mathtt{C}(u_1, \ldots, u_d) \tag{1}$$

On the other hand, when $A_{\mathtt{C}} = 0$ on $[0,1]^d$, then the copula is said to be *singular* and $\frac{\partial^d}{\partial u_1, \ldots \partial u_d} \mathtt{C}(u_1, \ldots u_d)$ in equation (1) is equal to 0 almost everywhere[1] in $[0,1]^d$, and the copula density does not exist [see, e.g., 12].

Note that when a copula has both an absolutely continuous component and a singular component, then its joint PDF cannot be easily obtained and it is sometimes undefined.

In this point, it is convenient to make a bold statement of the misuse of the copula density. Many authors, in geotechnical engineering, happily assume that every copula has a density which, as shown before, is not always the case. This is done because those authors avoid employing Stieltjes integrals, probably due to its unawareness, when in fact the use of these integrals simplifies the notation and the subsequent demonstrations.

2.3 Sklar's theorem

Sklar's theorem states that a continuous multivariate probability function can be expressed through its one dimensional marginal distributions and dependence structure. This dependence structure is, in fact, a copula function.

Let us consider a multivariate distribution function $F_{X_1, X_2, \ldots X_d}$ with marginal distributions F_{X_i} for $i = 1, \ldots, d$. Then, the Sklar's theorem dictates that there exists a copula $\mathtt{C}$ such that:

$$F_{X_1, X_2, \ldots, X_d}(x) = \mathtt{C}\big(F_{X_1}(x_1), F_{X_2}(x_2), \ldots, F_{X_d}(x_d)\big) \tag{2}$$

for all $x := [x_1, x_2, \ldots x_d] \in \mathbb{R}^d$. This copula is *unique* if all marginal distributions F_{X_i} are continuous; otherwise, the copula $\mathtt{C}$ is uniquely defined on the Cartesian product of the range of each marginal distribution, that is $\mathrm{range}(F_1) \times \mathrm{range}(F_2) \times \cdots \times \mathrm{range}(F_d)$.

Analogously, by knowing a copula $\mathtt{C}$ and a group of one dimensional CDFs F_{X_i}, then the d-dimensional distribution function $F_{X_1, X_2, \ldots, X_d}$ can be defined by the equation (2). The proof of Sklar's theorem for $d = 2$ is shown in Sklar [4], while the proof for $d > 2$ is stated in Sklar [14].

Note that equation (2) can be rewritten in order to obtain the inverse relation, which is given, for any $u := [u_1, u_2, \ldots, u_d] \in [0,1]^d$, by

$$\mathtt{C}(u) = F_{X_1, X_2, \ldots, X_d}\Big(F_{X_1}^{(-1)}(u_1), F_{X_2}^{(-1)}(u_2), \ldots, F_{X_d}^{(-1)}(u_d)\Big)$$

where $F_{X_1}^{(-1)}$ is the quasi-inverse of F_{X_i}, for $i = 1, 2, \ldots, d$.

[1] It is said in a probabilistic context that a property holds *almost everywhere* if it holds for all elements in a set except a subset whose measure is zero.

Furthermore, if the copula is absolutely continuous, the copula density exists and the d-dimensional PDF of the random variables $X_1, X_2, \cdots, X_d$ can be expressed as:

$$f_{X_1,X_2,\ldots,X_d}(x) = \mathtt{c}\big(F_{X_1}(x_1), F_{X_2}(x_2), \ldots, F_{X_d}(x_d)\big) \prod_{i=1}^{d} f_{X_i}(x_i) \quad (3)$$

where $\mathtt{c}$ is given by the equation (1), and f_{X_i} are the marginal density functions.

2.4 The Frechét-Hoeffding bounds and the independence copula

Based on the previously described properties, it can be demonstrated that any copula, regardless of its family or even its particular properties, is bounded by the inequality:

$$W_d(u) \leq \mathtt{C}(u) \leq M_d(u),$$

the so-called *Frechét-Hoeffding bounds*; here the lower limit W_d represents *countermonotonicity*, i.e. perfect negative dependence, and the upper limit M_d represents *comonotonicity*, i.e. perfect positive dependence.

In a bivariate case, W_2 and M_2 are expressed as [see, e.g., 11, p. 11]:

$$W_2(u_1, u_2) = \max(u_1 + u_2 - 1, 0)$$
$$M_2(u_1, u_2) = \min(u_1, u_2)$$

and in the d-dimensional case they are given by:

$$W_d(u) = \max(u_1 + u_2 + \cdots + u_d - d + 1, 0)$$
$$M_d(u) = \min(u_1, u_2, \ldots, u_d).$$

Another structure of dependence worth mentioning is in fact the one that represents independence. From probability theory, it is well known that the probability of occurrence of independent events is equal to the product of the probability of each particular event. Thus, by extending this concept to copula theory, independence is represented as:

$$\Pi_d(u) = \prod_{i=1}^{d} u_i = u_1 u_2 \cdots u_d.$$

Note that both M_d and Π_d are copulas for two or higher dimensions, and these are known as the *comonotonic* and the *product/independence copula*, respectively. On the other hand, W_d can only be considered a copula in a two-dimensional case, the so-called *countermonotonic copula* W_2, since for higher dimensions it does not fulfill the requirements to be a copula function. Nonetheless, the Fréchet-Hoeffding lower bound still holds for $d \geq 3$, even if the countermonotonic copula W_d does not exist [see 11, p. 47]. Figure 2.4 presents the *Fréchet-Hoeffding bounds* in the two-dimensional case, as well as the independence copula.

2.5 Some copula properties

In this section, some properties of copulas are discussed, in particular the range of dependence values, radial symmetry and tail dependence. These copula properties are useful when evaluating the suitability of a copula for modeling different dependence structures between random variables.

2.5.1 Range of dependence values

As stated before, any copula is bounded by the Fréchet-Hoeffding inequality. However, not all copulas have the ability to approach both limits, but are restricted to a narrower range. In other words, every copula is capable of modeling a certain spectrum of dependence within the Fréchet-Hoeffding bounds.

Copula functions with the property of approaching both countermonotonicity and comonotonicity limits are known as *comprehensive copulas* and are especially useful to model random variables whose dependence may vary between strongly positive or negative; some examples of comprehensive copulas are the Gaussian, Student-t, Plackett or Frank copulas. On the other hand, some other copulas can only approach one of the two limits; for instance, Gumbel or Joe copula can approach the comonotonic copula M_d, meanwhile the bivariate Nelsen No. 16 copula can approach the countercomonotonic copula W_2. Finally, some other copulas are not capable approaching any of the Fréchet-Hoeffding bounds, such as the Ali-Mikhail-Haq copula, which is restricted to a narrower range of dependence.

It is important to take this property into account when evaluating the suitability of a copula to model certain dependence structures. For example, if the dependence between random variables is known to be strongly negative, the Gumbel, Joe, or Ali-Mikhail-Haq copulas would not be a good choice for modeling this dependence. More details of these characteristics for certain copulas of interest will be presented in section 4.

2.5.2 Radial symmetry

It is said that a random vector $X \in \mathbb{R}^d$ (or its corresponding distribution) is radially symmetric about $a \in \mathbb{R}^d$ if

$$X - a \overset{d}{=} a - X,$$

where $\overset{d}{=}$ stands for the *equality in distribution*[2].

Analogously, a copula $\mathtt{C}$ is said to be radially symmetric if the random vector U, that follows the distribution $\mathtt{C}$, has the property [5, chapter 5]:

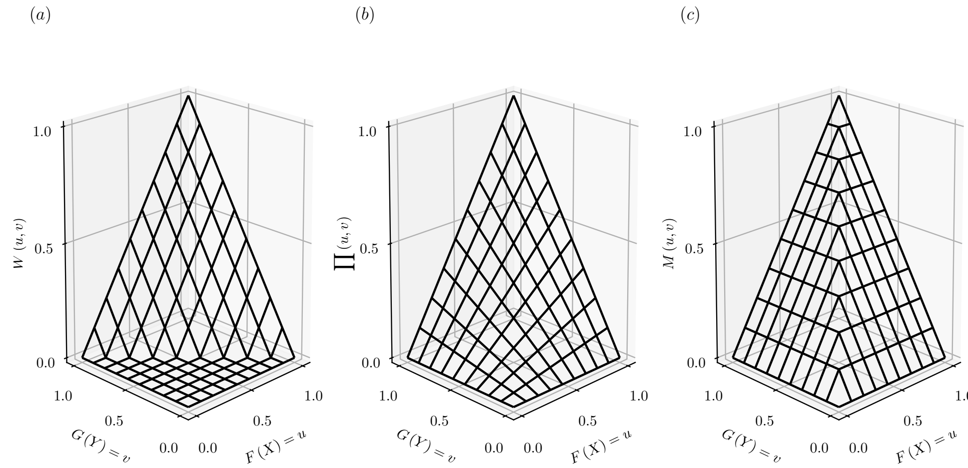
$$U \overset{d}{=} 1 - U.$$

Examples of radially symmetric copulas are Gaussian and Student-t copulas, while examples of non-radially symmetric copulas are Gumbel and Clayton copulas.

There are other concepts of symmetry in copula theory, such as the *marginal symmetry* or the *joint symmetry*; however, these

[2] If the sample space is a subset of the real line $\mathbb{R}$, then two random variables X and Y are said to be equal in distribution $X \overset{d}{=} Y$ when they have the same CDF, that is $P(X \leq x) = P(Y \leq x)$ for all x.

Figure 1 Fréchet-Hoeffding bounds and independent copula in two dimensions. (a) countermonotonic copula W_2, (b) independence copula Π, (c) comonotonic copula M_2.



properties are not exposed in this document; readers interested in these concepts are referred to Nelsen [11].

2.5.3 Tail dependence

Let us consider two random variables, X and Y with marginal CDFs F_X and F_Y respectively, associated through a copula $\mathtt{C}$. The coefficient of upper tail dependence λ_U and the coefficient of lower tail dependence λ_L are defined as follows [11]:

$$\lambda_U = \lim_{q \to 1^-} P(F_Y(y) > q | F_X(x) > q) \tag{4}$$

$$\lambda_L = \lim_{q \to 0^+} P(F_Y(y) \leq q | F_X(x) \leq q). \tag{5}$$

Conceptually, *upper tail dependence coefficient* λ_U is the conditional probability that Y exceeds the quantile q given that X also exceeds the same quantile q, as the quantile approaches to 1. Analogously, *lower tail dependence coefficient* λ_L is the conditional probability that Y falls behind the quantile q given that X also is smaller then the same quantile q, as the quantile approaches 0.

Assuming that both X and Y are continuous, equations (4) and (5) can be expressed as [11]:

$$\lambda_U = \lim_{q \to 1^-} \frac{1 - 2q + \mathtt{C}(q,q)}{1-q} \tag{6}$$

$$\lambda_L = \lim_{q \to 0^+} \frac{\mathtt{C}(q,q)}{q}. \tag{7}$$

These equations show that the coefficients of tail dependence are sole properties of the copula, and not of the marginal distributions. Furthermore, the existence or absence of one of these coefficients does not imply the existence or absence of the other. In other words, a copula $\mathtt{C}$ may have both upper and lower tail dependence, only one of those coefficients or none.

It is important to evaluate the dependence structure in the tails of the distributions since the failure domain is usually located there, and hence tails have a great impact on the probability of failure [15].

For this reason, coefficients of tail dependence are key measures of dependence in copula theory. Coefficients of tail dependence aim to quantify the associativity between random variables in the upper-right or lower-left quadrant of the d-dimensional box; or in other words, when all random variables take extreme values simultaneously.

2.6 Empirical copulas

In geotechnical engineering practice, it is common to have a limited sample available. With that sample one may try to infer the underlying copula. As the sample size increases the inferred copula will converge to the true copula. However, instead of inferring a copula, it is also possible to construct an *empirical copula* as follows.

Let us consider a set of N observations $\left\{\left(x_1^i, x_2^i, \ldots x_d^i\right) : i = 1, \ldots, N\right\}$ from a d-dimensional random vector $X = (X_1, X_2, \ldots, X_d)$. The pseudo-copula samples can be defined as follows:

$$\hat{U}_k^i = \frac{R_k^i}{N}$$

where R_k^i is the rank of the observation x_k^i:

$$R_k^i = \sum_{j=1}^{n} \mathbb{I}\left(x_k^j \leq x_k^i\right).$$

Hence, the corresponding empirical copula is then defined as

$$\mathtt{C}_N(u_1, \ldots, u_d) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\hat{U}_1^i \leq u_1\right] \times \cdots \times \mathbb{I}\left[\hat{U}_d^i \leq u_d\right]. \tag{8}$$

In this sense, the empirical copula can be understood as the empirical distribution of the rank transformed data. As the sample size increases, the empirical copula will approach to the true underlying copula of the multivariate data.

2.7 Vine copulas

Constructing copulas in high dimensions is not a trivial problem. Although there is a large number of copulas in the literature, most of them are bivariate or applicable in low-dimensions. In practice, the group of copulas in high dimensions is rather small and their properties are limited. There are mainly two drawbacks when trying to implement copulas in high dimensions:

1. There is no analytical solution for the extension to high-dimensions of the copula function, or their extension may become unfeasible.
2. The dependence structure becomes too restrictive and it is not possible to evaluate different dependence structures between pairs of variables.

For example, if one wants to build a geotechnical model with five random variables, there would have few copula options to choose from and one would have to assign the same dependence structure to all variables, which may be different from reality.

Fortunately, in recent years this problem has been circumvented by the so called *vine copula theory* [16]. In this section, we briefly introduce it mostly based on Aas et al. [17] and Czado [18].

Vine copula theory states that it is possible to express a multivariate PDF as a product of pair-copula densities and marginal densities. Vine copula theory employs bivariate copulas as building blocks for elaborating higher-dimensional distributions, where the dependence is determined by these bivariate copulas in a nested set of trees. In this way, vine copulas allows to construct models in high dimensions, in which it is possible to evaluate different structures of dependence between each pair of variables.

To clarify these concepts, let us consider a vector of d random variables $X = [X_1, \ldots, X_d]$ distributed according to the PDF f. It is possible to decompose f as:

$$f(x_1, \ldots, x_d) = f(x_d) f(x_{d-1}|x_d) f(x_{d-2}|x_{d-1}, x_d) \cdots f(x_1|x_2, \cdots, x_d); \quad (9)$$

this decomposition is unique up to a relabeling of the variables [17].

Now, observe that if the copula $\mathtt{C}$ is absolutely continuos, and hence the copula density $\mathtt{c}$ exists, equation (3) could be expressed for a bivariate case as:

$$f(x_1, x_2) = \mathtt{c}_{12}(F_1(x_1), F_2(x_2)) f(x_1) f(x_2)$$

and hence the conditional densities as:

$$f(x_1|x_2) = \mathtt{c}_{12}(F_1(x_1), F_2(x_2)) f_1(x_1). \quad (10)$$

The copula conditional density of equation (10) can be employed in equation (9) replacing each conditional term. In this way, a joint PDF can be decomposed into pair-copulas times conditional marginal densities. For instance, one possible decomposition for a three-dimensional PDF would be:

$$f(x_1, x_2, x_3) = f_1(x_1) f_{2|1}(x_2|x_1) f_{3|12}(x_3|x_1, x_2)$$

where

$$\begin{aligned} f_{2|1}(x_2|x_1) &= \mathtt{c}_{12}(F_1(x_1), F_2(x_2)) f_2(x_2) \\ f_{3|12}(x_3|x_1, x_2) &= \mathtt{c}_{13|2}\big(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)\big) f_{3|2}(x_3|x_2) \\ f_{3|2}(x_3|x_2) &= \mathtt{c}_{23}(F_2(x_2), F_3(x_3)) f_3(x_3) \end{aligned}$$

and hence:

$$\begin{aligned} f(x_1, x_2, x_3) &= f_1(x_1) f_2(x_2) f_3(x_3) \\ &\times \mathtt{c}_{12}(F_1(x_1), F_2(x_2)) \mathtt{c}_{23}(F_2(x_2), F_3(x_3)) \\ &\qquad \times \mathtt{c}_{13|2}\big(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)\big). \end{aligned}$$

Nonetheless, as stated before this decomposition is not unique, and as the number of variables in the model increases, the number of possibilities of pair-copula constructions also grows. For instance, a model of three variables has 6 possibilities of decomposition, a model of four variables has 24 possibilities of decomposition, and a model of five variables has 240 possibilities of decomposition [17].

For the sake of organizing all these options of decomposition, Bedford and Cooke [19] introduced a graphical tool called *the regular vine*. A regular vine is a graphical aid for labeling bidimensional constraints in high-dimensional PDFs. When regular vines are combined with bivariate copulas, vine copulas are established.

A regular vine of d variables can be seen as a nested set of connected trees, in which the edges of the tree T_j are the nodes of the tree T_{j+1}, and the edges of the tree T_{j+1} are the nodes of the tree T_{j+2}, and so on. Furthermore, two edges in the tree T_j are joined in the tree T_{j+1} if and only if the edges share a common node. In a regular vine copula, the nodes in the first tree are the random variables, and from then on each edge and node is a conditional copula density.

Regular vines are still a quite general category that embraces several classes of pair-copula decompositions. Among all the variants of the regular vines, it is worth mentioning two special cases: the *D-vine* and the *canonical vine*. Each one of these two classes of regular vines has a specific way of describing the decomposition of a density function.

In a D-vine, each tree is a path. Therefore, a D-vine of d variables would have T_j trees, with $j = 1, \ldots, d-1$, and each tree would have $d+1-j$ nodes and $d-j$ edges. Note that each edge corresponds to a pair-copula density and each edge label corresponds to the subscript of the pair copula density. Figure 2.7 presents an example of a D-vine with 5 variables.

From figure 2.7 the decomposition of the 5-dimensional probability model can be expressed as:

$$f_{12345} = f_1 f_2 f_3 f_4 f_5 \; \mathtt{c}_{12} \mathtt{c}_{23} \mathtt{c}_{34} \mathtt{c}_{45} \; \mathtt{c}_{13|2}$$

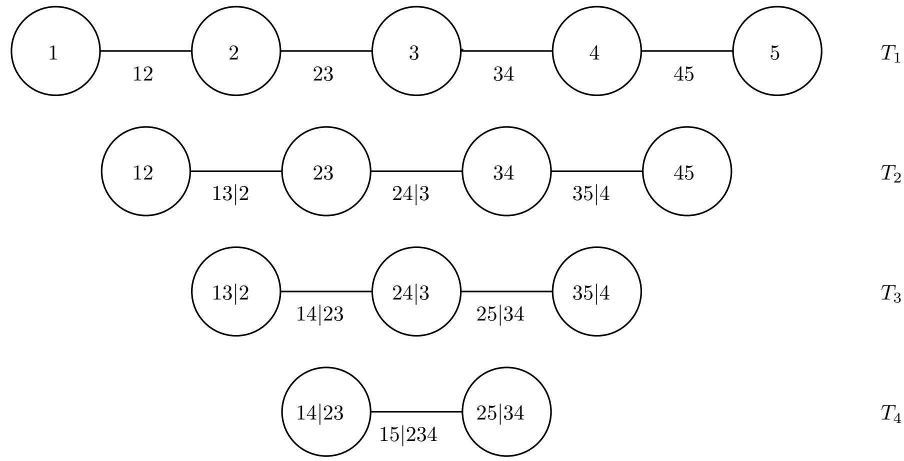


Figure 2 Example of a D-vine with 5 variables, 4 trees and 10 edges. Each edge (represented as a line) is associated with a pair-copula conditional density.

$$c_{24|3}c_{35|4}\ c_{14|23}c_{25|34}\ c_{15|234}$$

The general equation that represents a d-dimensional PDF through a D-vine is:

$$f(x_1,\ldots,x_d)=\prod_{k=1}^{d}f(x_k)\times\prod_{j=1}^{d-1}\prod_{i=1}^{d-j}c_{i,i+j|i+1,\ldots,i+j-1}\left(F\left(x_i|x_{i+1},\ldots,x_{i+j-1},\right)F\left(x_{i+j}|x_{i+1},\ldots,x_{i+j-1}\right)\right).$$

In a canonical vine, each tree T_j has an unique central node that connects to the other nodes through $d-j$ edges. This type of structure is advantageous when one knows that a particular variable is the one that governs the interactions in the dataset. For instance, it is typical in geotechnical engineering to relate the $(N_1)_{60}$ of the SPT to multiple parameters and variables of soils; then, it would be desirable to recreate a multivariate model where $(N_1)_{60}$ is at the root of the canonical vine. Figure 3 presents an example of a canonical vine with 5 variables.

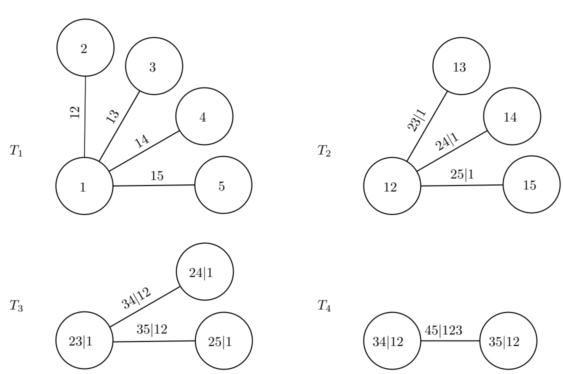


Figure 3 Example of a canonical vine with 5 variables, 4 trees and 10 edges. Each edge is associated with a pair-copula conditional density.

From figure 3 the decomposition of the 5-dimensional probability model can be expressed as:

$$f_{12345}=f_1f_2f_3f_4f_5\ c_{12}c_{13}c_{14}c_{15}\ c_{23|1}c_{24|1}\ c_{25|1}\ c_{34|12}c_{35|12}\ c_{45|123}$$

The general equation in which a d-dimensional probability density function is expressed through a canonical vine is given by:

$$f(x_1,\ldots,x_d)=\prod_{k=1}^{d}f(x_k)\times\prod_{j=1}^{d-1}\prod_{i=1}^{d-j}c_{j,j+i|1,\ldots,j-1}\left(F\left(x_j|x_1,\ldots,x_{j-1}\right),F\left(x_{j+i}|x_1,\ldots,x_{j-1}\right)\right)$$

Constructing vine copulas from a dataset encompass three general steps. First, the structure of the vine must be specified. That is to say, the type of the vine and the interaction of the nodes (and variables) on each tree. Second, a bivariate copula function must be established for each edge on each tree. A wide range of bivariate parametric copulas exist in the literature and can be employed. Third, each copula must be adjusted to the available data. Several methodologies exists in the literature for this last purpose.

Some of these concepts will be discussed in more detail in this document. In particular, section 4.3 presents several families of copulas that are popular in geotechnics; section 4.4 explains different methodologies for adjusting copulas to data and the issue of defining the best copula for a dataset is treated in section 4.5.

3 Some remarks about dependence in geotechnics

In order to understand the essence of copulas and their applicability in geotechnical engineering, it is necessary to clarify what dependence and correlation are and how they are related to multivariate CDFs. Although these two concepts are sometimes assumed as synonyms, from a statistical perspective they have different meanings.

On the one hand, dependence can be better approached from the concept of statistical independence: two random variables X and Y are said to be *statistically independent* if the occurrence of one event X does not have implications on the CDF of another event Y. This can be mathematically written through conditional probabilities as $P(X|Y) = P(X)$ or $P(Y|X) = P(Y)$; thus, two random variables are considered *dependent* if the conditions to be statistically independent are not satisfied.

On the other hand, *correlation* is a statistical measure of association between two random variables. It must bear in mind that the term correlation is sometimes understood as a measure of linearity between two random variables, such as the Pearson's rho correlation coefficient does [see, e.g., 20], although this term is also applicable to other non-linear measures of dependence, such as the rank measures that will be studied in section 4.2.

Dependence is a more general concept than correlation, and in fact, it encompasses all types of statistical correlations. Correlation implies dependence, but dependence does not necessarily implies correlation, e.g. a Pearson's rho correlation coefficient equal to zero does not mean statistical independence between two random variables, since there could be other structures of dependence different from linearity.

Following these concepts, there are many variables in geotechnics that can be considered dependent, such as the shear strength parameters of the Mohr-Coulomb failure criterion, Atterberg limits, parameters of the soil-water characteristic curves, foundation settlement parameters, among others. Geotechnical literature is full of correlations and equations to predict values of some parameters based on some others that are easier to obtain. For instance, the $(N_1)_{60}$ of SPT is commonly related to many other parameters such as the unconfined compressive strength q_u, cone tip resistance q_c, soil friction angle, among others. Readers interested in some of these correlations are referred, for instance, to Ameratunga et al. [21].

Additionally, it is found that the most used measure of correlation in geotechnical engineering corresponds to the Pearson's rho coefficient; this is expected given its popularity, simplicity, and that it is the canonical measure of dependence of the widely used multivariate normal distribution. This coefficient and its implications will be further studied in section 4.2.

Now, starting from the concept of dependence, it is necessary to construct multivariate CDFs that indicate the probability of occurrence of two or more dependent events. These functions are essential to simulate random values with certain dependence structures or to carry out reliability analysis and compute probabilities of failure of geotechnical models. However, in practice, there are a few multivariate CDFs that are easy to construct and extend to high dimensions, which narrows the range of possibilities when representing dependence through a CDF.

Furthermore, it is unfortunately usual in geotechnical engineering practice to have a reduced amount of data due to quite limited budgets or lack of time. With this scarce information at hand, it is only possible to define marginal distributions, i.e. univariate distribution functions for each random variable, and a correlation matrix, which is commonly constructed through Pearson's rho coefficients among random variables. This condition is referred to as *incomplete probability information*, and a basic introduction to its characteristics is found in Der Kiureghian and Liu [22].

Under these circumstances, the multivariate normal distribution is commonly adopted, without validation, for representing probabilities of the joint occurrence of the dependent random variables. It is well known that this distribution has many advantages since it has been widely studied, its extension to high dimensions is straightforward, it only requires the definition of a covariance matrix Σ based on Pearson's rho coefficients, and in general, its application is quite simple. The d-dimensional normal PDF is given by:

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^d \det\Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{11}$$

where X is a random vector in $\mathbb{R}^d$, $\mu \in \mathbb{R}^d$ is the mean vector and $\Sigma \in \mathbb{R}^{d\times d}$ is the covariance matrix constructed through Pearson's rho correlation coefficients.

However, equation (11) is too restrictive since all random variables are assumed to be normally distributed, and dependence is limited uniquely to a Gaussian type. This results in drawbacks that appear when modeling joint occurrences of some geotechnical random variables, such as shear strength parameters, where physically both cohesion and friction angle cannot be negative.

In consequence, there is a need of exploring other alternatives in this regard. Some early works, such as Der Kiureghian and Liu [22], studied the issue of the representation of random vectors and the computations of the probabilities of failure under incomplete probability information. However, although there are several possibilities for this purpose, the truth is that very few are applicable in practice since the majority are restricted only to a bivariate case and/or only are capable of representing small correlations between random variables. One of the most widely used alternatives, before copula theory became popular, was the Nataf transformation, which will be introduced in the next.

3.1 The Nataf transformation

The Nataf transformation [3] is an approach that appears as a solution to the problem of modeling dependence when probability information is reduced to marginal distributions of all random variables and a linear correlation matrix ρ.

Following Liu and Der Kiureghian [23] and Li et al. [24], the Nataf transformation can be defined as a transformation of random variables from the original/physical space to a space of standard normal variables. In this sense, a d-dimensional vector of dependent random variables X, with marginal CDFs $F_{X_i}(x_i)$, is transformed into a vector of random variables Z that follows a normal probability function $\mathcal{N}(0,\rho_0)$ with linear correlation matrix ρ_0 using the transformation $Z_i = \Phi^{-1}[F_{X_i}(x_i)]$ for $i = 1,\ldots,d$; here Φ denotes the standard normal CDF.

Additionally, based on the Nataf transformation, the joint PDF of random vector X in the original/physical space can be expressed as:

$$f_X(x) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_d}(x_d) \frac{\phi_d(z;\, 0, \rho_0)}{\phi(z_1)\phi(z_2)\cdots\phi(z_d)} \quad (12)$$

where ϕ_d is the d-dimensional normal PDF with mean zero and correlation matrix ρ_0. In general, equation (12) is known as *the Nataf distribution*.

It must bear in mind that the Nataf transformation causes a *correlation distortion*; that is, the correlation coefficient $\rho_{0_{ij}}$, between the standard normal variables Z_i and Z_j, is not equal to its counterpart within variables X_i and X_j, denoted by ρ_{ij}. The correlation matrix ρ_0 can be obtained from ρ by applying the following equation [22]:

$$\rho_{ij} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(\frac{F_i^{-1}(\Phi(z_i)) - \mu_i}{\sigma_i} \right) \left(\frac{F_j^{-1}(\Phi(z_j)) - \mu_j}{\sigma_j} \right) \phi_{Z_i,Z_j}\left(z_i, z_j, \rho_{0_{ij}}\right) dz_i dz_j;$$

this equation can be solved iteratively through numerical methods. Nonetheless, some semi-empirical formulas for approximating ρ_0 from ρ, under different conditions, were presented by Der Kiureghian and Liu [22]. Also, Li et al. [24] presented a Gauss-Hermite-Quadrature method which leads to better approximations of ρ_0 than the semi-empirical formulas of Der Kiureghian and Liu [22].

As we can observe, with the use of the Nataf transformation we can couple dependent random variables with different marginal PDFs; this is is more flexible than the multivariate normal distribution of equation (11). Additionally, Nataf transformation integrates very well with popular reliability methods, such as the First Order Reliability Method (FORM) or the Monte Carlo simulation (MCS) [see, e.g., 24]. Due to this, the Nataf transformation has been the default method for many years to handle dependent random variables with different univariate distributions.

However, Nataf transformation has several implicit hypotheses, and the most notable is the one of the dependence structure. In recent years, copula theory has given new insights into dependence, and it has allowed to elucidate the hidden dependence hypothesis behind Nataf transformation. In this regard, it was demonstrated by Lebrun and Dutfoy [25], in the light of copula theory, that Nataf transformation is nothing but a particular case of the Gaussian copula.

Now, what are the implications of Nataf transformation being a particular case of Gaussian copula? In fact, in the light of copula theory, Gaussian copula is just one structure of dependence that belongs to the wide range of dependence structures than can be modeled through copulas. In this way, Gaussian copula, or Nataf transformation, is one of many options and perhaps it may not be the best structure of dependence for representing a dataset, leading thus to misleading models. This topic is further explored in subsequent chapters of this document.

It is worth mentioning that, in addition to the work of Lebrun and Dutfoy [25], a generalization of the Nataf transformation to all kind of elliptical copulas was exposed by Lebrun and Dutfoy [26]. In this sense, Nataf transformation was extended to represent any elliptical copula, not just a Gaussian copula. Besides this, the popular FORM/SORM were also extended to this generalized Nataf transformation.

The advantage of copulas is evident since, without them, there is only a narrow range of multivariate distribution functions in practice and those are too restrictive in terms of dependence. Additionally, copula theory is more general than the widely used Nataf transformation, and hence its range of possibilities is larger. Its multiple advantages have led copulas to be applied in many fields, being geotechnics one of them.

4 Construction of copulas for modeling dependence among geotechnical variables

There is a great variety of copulas in the literature, each with its own particular properties that make them suitable for representing different dependence conditions. However, considering this wide range of possibilities, the question arises as to which is the best copula to model the dependence between our variables of interest.

Therefore, this section will be dedicated to elucidate, from a practical perspective, how this question can be solved in the best possible way based on the available information. In other words, the concepts of the construction of copulas to model geotechnical parameters will be introduced in this section, based on the latest postulates of the state-of-the-art. Nonetheless, it is worth mentioning that the methodologies presented here are not a straitjacket and different variants can be applied depending on the particular conditions of the problem of interest. In this way, the content of this section is a guide that readers can adopt as a basis for their particular problems.

Five general steps are identified in the construction and definition of a copula function for the purposes of modeling geotechnical variables:

1. Define and fit the marginal PDF of each random variable involved in the model.
2. Estimate the strength of the dependence between the random variables.
3. Define a set of candidate copulas $\mathscr{C}$ that could fit the dependence characteristics of the random variables of interest.
4. Fit the candidate copulas in $\mathscr{C}$ to the available information.
5. Select the most suitable copula in $\mathscr{C}$ through some goodness-of-fit test.

Each of these steps will be covered in detail. In particular, section 4.1 provides some guidelines on the adjustment of random variables to one-dimensional PDFs; section 4.2 introduces the concept of correlation, and explains its role as a first approximation to the magnitude of dependence; section 4.3 introduces those copulas that, after reviewing the state-of-the-art, were identified as the most frequently used in geotechnics; section 4.4 explains how to fit copulas to the available data. Finally, section 4.5

introduces some practical techniques for the definition of the most suitable copula among a set of candidate copulas $\mathscr{C}$.

4.1 Selection of marginal distributions

There is a wide variety of univariate PDFs in the literature, both continuous and discrete (see e.g. Montgomery and Runger [27], Ang and Tang [20] or Kottegoda and Rosso [28]). In particular, we will be interested in continuous PDFs. Some of them have been quite popular in geotechnics, such as the normal, lognormal and uniform. Some others have also shown a great potential for modeling geotechnical parameters, such as the exponential, gamma, beta, Weibull or Gumbel.

In general, the process of fitting and selecting a marginal distribution is simple and straightforward to implement. Firstly, a set of candidate distributions must be proposed. Then, these distributions area fitted to the data set using, for example, standard techniques such as the method of moments or the maximum likelihood method [see, e.g., 20]. Subsequently, the best fit distribution will be selected and for this purpose different goodness of fit techniques can be employed. Among all the existing goodness of fit techniques, some widely used are worth mentioning, such as the Kolmogorov-Smirnov test, Cramér-von-Mises criterion, Anderson–Darling test, the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). These last two techniques will be discussed in Section 4.5.2.

4.2 Measures of correlation

Dependence is one of the most studied and important concepts in probabilistic and statistic analysis, and hence in reliability studies. Stochastic dependence has a great impact not only in the construction of a joint PDF but also in the estimation of the probability of failure [29, 11].

In practice, there is a wide range of degrees of dependence among geotechnical random variables. However, dependence in geotechnics is commonly characterized by a number that varies in the interval $[-1,1]$ and that evaluates its strength. These coefficients are commonly known as *measures of correlation* and although they are quite useful and easy to calculate, they serve only as a first approximation to the whole structure of dependence of two random variables.

There is a wide variety of correlation measures in the literature, each with its own aims, scopes, and particular properties [see, e.g., 30]. However, in this document, we are mainly interested in two categories: the linear correlation measures and the measures of concordance. On the other hand, linear correlation measures are important as they are the most popular in practice, and they are also the basis for elliptical distributions (a concept that will be discussed later), such as the multivariate Gaussian distribution. However, their use in non-elliptical distributions is misleading, erroneous and sometimes useless [31, 12]. On the other hand, the concordance measures are important since they are the most suitable correlation measures to be used in conjunction with copula theory.

The concepts of linear correlation and concordance will be discussed in more detail below. In particular, the Pearson's rho correlation coefficient will be introduced as a measure of linear correlation and the Spearman's rho and Kendall's tau coefficients will be introduced as measures of concordance.

4.2.1 Linear correlation

Linear relationship between two random variables, X and Y, can be expressed by the following equation:

$$E(Y|X=x) = \alpha + \beta x$$

where α is the intercept and β the slope of the straight line that relates both random variables. This equation states that by having X, one could predict the values of Y by employing a direct linear relationship. Nonetheless, two random variables are seldom perfectly associated in a linear way. Conversely, the straight line that relates both variables must be fitted to the available information, and even so the dispersion of the data may not allow a good adjustment. Thus, it is necessary to evaluate the suitability of the linear relationship between two random variables and for this purpose the coefficients of linear correlation are employed. According to Embrechts et al. [12] and Ang and Tang [20], linear correlation can be defined as a measure of suitability of linear dependence. In particular, *Pearson's rho correlation coefficient* is introduced below.

Pearson's rho

This is the most popular used measure of dependence, given its simplicity and its intrinsic relationship with the multivariate Gaussian PDF. Pearson's rho correlation coefficient is a measure of linear dependence between two random variables, X and Y, as:

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}. \tag{13}$$

Note that an extension to more than two dimensions ($d > 2$) is straightforward, since the correlation is evaluated pairwise so that a $d \times d$ *correlation matrix* is formed. For example, given three random variables, X_1, X_2 and X_3, the correlation matrix is expressed as:

$$\rho = \begin{bmatrix} 1 & \rho(X_1,X_2) & \rho(X_1,X_3) \\ \rho(X_2,X_1) & 1 & \rho(X_2,X_3) \\ \rho(X_3,X_1) & \rho(X_3,X_2) & 1 \end{bmatrix}.$$

It can be shown from equation (13) that $\rho(X_1,X_2) \in [-1,+1]$; here, -1 or $+1$ would mean a perfect linear relationship between the given two variables, either positive or negative and 0 would mean linear uncorrelatedness. It is worth noting that linear uncorrelatedness does not mean independence, since two variables may have no linear relationship but rather other types of relationships [see, e.g., 20].

Nonetheless, equation (13) expresses that Pearson's rho correlation between two random variables depends not only on the covariance of these variables, but also on their variance and consequently on their marginal distributions. In this sense, we can extract two important remarks:

1. The range of values that $\rho(X_1, X_2)$ can adopt will be bounded between ρ_{min} and ρ_{max} (i.e., $[\rho_{min}, \rho_{max}] \subset [-1, +1]$), where ρ_{min} and ρ_{max} are functions of the variance of the marginal distributions.
2. Pearson's correlation coefficient ρ depends on both, joint and marginal distributions of the involved random variables. This behavior is against the nature of copula theory, where marginal distributions and dependence are isolated.

Inasmuch as Pearson's correlation coefficient is only invariant in strictly increasing linear transformations, but not in strictly increasing transformations, it is necessary to explore other measures of dependence that better agree with the nature of the copulas; the measures of concordance are especially suitable for this purpose.

4.2.2 Concordance

Before introducing the two main rank correlation coefficients, namely *Spearman's rho* and *Kendall's tau*, it is important to explore the concept of concordance.

Following Nelsen [11], let us assume two independent observations (x_i, y_i) and (x_j, y_j) from a vector of continuous random variables (X, Y). The pair of observations are said to be *concordant* if $x_i < x_j$ whenever $y_i < y_j$ or if $x_i > x_j$ whenever $y_i > y_j$. In the same way, these two observations are said to be *discordant* if $x_i < x_j$ whenever $y_i > y_j$ or if $x_i > x_j$ whenever $y_i < y_j$. Alternatively, it could be formulated that two observations (x_i, y_i) and (x_j, y_j) are concordant if $(x_i - x_j)(y_i - y_j) > 0$, or discordant if $(x_i - x_j)(y_i - y_j) < 0$. In this sense, concordance is a measure of associativity between large or small values of two random variables.

Rank correlations are native measures of concordance for two random variables [5]. Essentially, a rank correlation is a scalar measure that is based on the ranks of the data instead of the actual numerical values. From copula theory viewpoint, rank correlation coefficients only depend on the copula and not on the marginal distributions, which contrasts with Pearson's rho correlation coefficient that depends on both. This fact makes rank correlations the natural and most appropriate measure of dependence for copulas [5].

With this context, the two well-known nonparametric measures of dependence, namely *Spearman's rho rank correlation* and *Kendall's tau rank correlation*, can be defined as follows.

Spearman's rho

Informally, Spearman's rho coefficient (ρ_s) could be defined as the linear correlation of the transformed random variables. In this way, given two random variables X and Y,with CDFs F_X and F_Y, respectively, Spearman's rho coefficient is defined as:

$$\rho_s(X, Y) = \rho(F_X(x), F_Y(y)).$$

When the probability distribution of the random variables is not certainty known, it is convenient to define Spearman's rho coefficient in terms of the ranks of the data. These ranks, denoted as R_i and S_i, result from the ordering of the sample of the two random variables, X and Y, that is, $R_i = \text{rank}(X_i)$ and $S_i = \text{rank}(Y_i)$ [see 20, for details]. Then, Spearman's rho can be, alternatively, defined as:

$$\rho_s = \frac{\sum_{i=1}^{n} (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n} (R_i - \bar{R})^2 \sum_{i=1}^{n} (S_i - \bar{S})^2}}$$

where

$$\bar{R} = \bar{S} = \frac{1}{n} \sum_{i=1}^{n} R_i = \frac{1}{n} \sum_{i=1}^{n} S_i = \frac{n+1}{2}.$$

Kendall's tau

Conceptually, Kendall's tau coefficient (τ) is defined as the probability of concordance minus the probability of discordance

$$\tau(X, Y) = P\big((X - \tilde{X})(Y - \tilde{Y}) > 0\big) - P\big((X - \tilde{X})(Y - \tilde{Y}) < 0\big)$$

where $\tilde{X}$ and $\tilde{Y}$ are an independent copies of X and Y, respectively, i.e, a second version with the same distribution but independent of the first.

Analogously, a sample version of Kendall's tau coefficient is expressed in the following equation:

$$\tau_n = \frac{P_n - Q_n}{\binom{n}{2}} = \frac{4}{n(n-1)} P_n - 1$$

where P_n is the number of concordant pairs and Q_n the number of discordant pairs of the sample.

Both Spearman's rho and Kendall's tau, as well as Pearson's rho correlation coefficient, are easily calculable in two dimensions. Furthermore, extensions of these coefficients to more than two dimensions are straightforward, since only pairwise evaluations of the variables are required to form a correlation matrix, as described above for the correlation matrix of Pearson's rho coefficients.

Rank measures of dependence, such as Spearman's rho and Kendall's tau, have several properties that make them especially suitable for their use in copula theory. For instance:

1. Both Spearman's rho and Kendall's tau are symmetric and take values within $[-1, 1]$, regardless the marginal distributions.
2. A value equal to zero in either of these two rank coefficients is a stronger indicator of independence than Pearson's rho coefficient; however, such hypothesis must be corroborated with other tests.
3. Rank correlation values of $+1$ or -1 indicate comonotonicity and countermonotonicity respectively [see, e.g., 13].
4. Both rank correlations do not depend on marginal distributions, but rather solely on the adopted copula.
5. Both rank coefficients, unlike linear correlation, are invariant under strictly increasing transformations and are also less sensitive to outliers.

So, rank measures, either Spearman's rho or Kendall's tau, are the most suitable and natural measures to be used in the representation of the strength of dependence between two random variables, in copula theory.

4.3 Popular families of copulas

A great variety of copulas could be found in the literature, each one with its properties and dependence structure. Also, it is common to group these copulas into some families, taking into account some similarities in their properties or in how these copulas are constructed. Particularly, in the geotechnical engineering framework, and after a review of the state-of-the-art, three families of copulas are mainly used namely Elliptical, Archimedean, and Plackett copulas.

In this section, we will review the aforementioned three families of copulas, while providing some theoretical background and explanation of their properties. For an in deep explanation, the reader is referred for example, to Nelsen [11], Joe [32] or Embrechts et al. [12].

There exist other families of copulas widely used and quite studied in several fields, such as the Farlie-Gumbel-Morgenstern (FGM), Ali-Mikhail-Haq (AMH) or Marshall-Olkin copulas. However, their utility has been rather scarce in geotechnics, mainly due to the reduced range of degrees of dependence that they are able to handle.

4.3.1 Elliptical copulas

Elliptical copulas are associated to elliptical distributions as a consequence of the *Sklar's theorem*. The shape of these distributions is symmetric and from here stems their name, since the isocurves of their PDFs are ellipse-like forms in $\mathbb{R}^n$ (see e.g. Figure 4.3.1).

An elliptical distribution, which is denoted as $E_d(\mu,\Sigma)$ in $\mathbb{R}^d$, mainly consists of two parameters: a *localization vector* $\mu \in \mathbb{R}^d$ and a *scatter matrix* $\Sigma \in \mathbb{R}^{d\times d}$ that is symmetric and positive semi-definite.

These distributions/copulas are radially symmetric around their respective localization vector μ, and their scatter is represented by the scatter matrix Σ. Another important remark about these distributions is that the linear correlation coefficient is their natural measure of dependence, so for a random vector, each element ρ_{ij} of the correlation matrix ρ can be defined as:

$$\rho_{ij} = \rho(X_i,X_j) = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}.$$

Disadvantages associated to linear correlation have already been studied before in this document. Furthermore, elliptical multivariate distributions, and the definition of their correlation matrix ρ, require all marginal distributions to be elliptical [12]. Nonetheless, these two pitfalls are solved using copulas.

In order to apply robust linear correlation estimators, rank measures of dependence have to be used. Specifically, the relation between Kendall's tau and Pearson's rho is suitable for this purpose. In this way, copula parameters of dependence of elliptical copulas, and hence their matrix of correlation, can be computed through the respective Kendall's tau coefficients between pairs of random variables. This relation is given by the following equation [33]:

$$\tau(X_i,X_j) = \frac{2}{\pi}\arcsin(\rho_{ij}) \tag{14}$$

Note that if $i = j$ then $\rho_{ij} = 1$.

With this brief introduction to elliptical distributions, we are now in a position to present two of the most used elliptical copulas in geotechnical engineering, namely the Gaussian copula and the Student-*t* copula. Readers interested in more details about theory of elliptical distributions and elliptical copulas can find a good exposure of this topic in Embrechts et al. [12].

Bivariate equations of both Gaussian copula and Student-*t* copula are presented in table 1. Both copulas are comprehensive copulas, i.e. they both can approach to countermonotonicity W_2 and comonotonicity M_d. Furthermore, equation (14) holds for these two copulas as a way of fitting them to a given Kendall's tau. Extending these two copulas to more than two-dimensions is straightforward, and this fact is one of the reasons of their popularity.

Although the copula dependence parameter of both copulas can be fitted by equation (14), Student-*t* copula has an additional parameter of adjustment, namely degrees of freedom ν. While Gaussian copula is a one-parameter copula, the Student-*t* copula is a two-parameter copula. The number of degrees of freedom of a Student-*t* copula controls the strength of both tail dependence coefficients, i.e. λ_u and λ_l (see equations (6) and (7)), as stated in the equations presented in table 2.

Note that as a consequence of the radial symmetry property of elliptical copulas, both coefficients of tail dependence of the Student-*t* copula are equal. Additionally, as ν increases, the tail dependence of the Student-*t* copula becomes weaker, and vice versa, as ν decreases, the tail dependence of the Student-*t* copula becomes stronger. In this way, when ν tends to infinity, both λ_u and λ_l approach to zero.

On the other hand, the Gaussian copula does not have tail dependence, i.e. $\lambda_u = \lambda_l = 0$. However, what is truly interesting here is that the Gaussian copula is a particular case of the Student-*t* copula. Student-*t* copula becomes a Gaussian copula when its tail dependence is equal to zero, i.e., when ν tends to infinity. Therefore, Student-*t* copula has more general dependence structure than Gaussian copula. Some authors, such as Embrechts et al. [12], support the use of the Student-*t* copula as an alternative to the commonly used Gaussian dependence structure.

The number of degrees of freedom of a Student-*t* copula can be estimated from data using fitting methods based on the concept of maximum-likelihood-estimation, like the ones presented in section 4.4. When no data is available and just the rank measure of dependence is known, some authors such as McNeil et al. [5] recommend to adopt a value of 4 for the degrees of freedom.

Algorithms for modeling both Gaussian and Student-*t* copulas can be found in Embrechts et al. [12] or McNeil et al. [5]. For illustrative purposes, Figure 4.3.1 shows the contours of two

Table 1 Bivariate copula function C and copula density function c for Gaussian and Student-*t* copulas

Copula	$\mathtt{C}(u,v;\ \theta)$	$\mathtt{c}(u,v;\ \theta)$
Gaussian	$\Phi_\theta\left(\Phi^{-1}(u),\Phi^{-1}(v)\right)$	$\frac{1}{\sqrt{1-\theta^2}}\exp\left[-\frac{\xi_1^2\theta^2-2\theta\xi_1\xi_2+\xi_2^2\theta^2}{2(1-\theta^2)}\right]$; $\xi_1=\Phi^{-1}(u);\ \xi_2=\Phi^{-1}(v)$
Student-*t*	$t_{\nu,\theta}\left(t_\nu^{-1}(u),t_\nu^{-1}(v)\right)$	$\frac{\Gamma\left(\frac{\nu+2}{2}\right)\Gamma\left(\frac{\nu}{2}\right)}{\Gamma^2\left(\frac{\nu+1}{2}\right)\left(1-\theta^2\right)^{1/2}}\frac{\left(1+\frac{\left(T_\nu^{-1}(u)\right)^2-2\theta T_\nu^{-1}(u)T_\nu^{-1}(v)+\left(T_\nu^{-1}(v)\right)^2}{\nu\left(1-\theta^2\right)}\right)^{-\frac{\nu+2}{2}}}{\left(1+\frac{\left(T_\nu^{-1}(u)\right)^2}{\nu}\right)^{-\frac{\nu+1}{2}}\left(1+\frac{\left(T_\nu^{-1}(v)\right)^2}{\nu}\right)^{-\frac{\nu+1}{2}}}$

$\Gamma =$ Gamma function.

$\nu =$ degrees of freedom.

$\theta =$ Pearson's rho correlation coefficient.

$t_{\nu,\theta} =$ univariate Student's PDF, with ν degrees of freedom and θ Pearson's rho correlation coefficient.

$T_{\nu+1} =$ univariate Student's t CDF, with $\nu+1$ degrees of freedom

$\Phi_\theta =$ bivariate Gaussian CDF, with zero mean vector and θ Pearson's rho correlation coefficient.

$\Phi^{-1} =$ inverse of a standard Gaussian CDF.

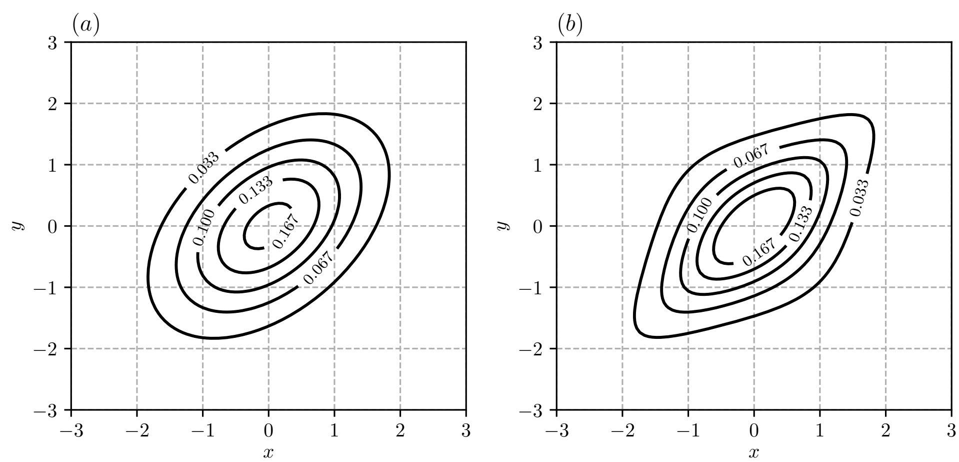


Figure 4 Contour plots of two bivariate PDFs whose copulas are elliptical and their marginals are standard normal PDFs. Both copulas have a Kendall's tau $\tau = 0.3$. (a) Gaussian copula; (b) Student-*t* copula ($\nu = 3$)

Table 2 Tail dependence of Gaussian and Student-*t* copulas

Copula	λ_l	λ_u
Gaussian	0	0
Student-*t*	$2T_{\nu+1}\left(-\frac{\sqrt{\nu+1}\sqrt{1-\theta}}{\sqrt{1+\theta}}\right)$	$2T_{\nu+1}\left(-\frac{\sqrt{\nu+1}\sqrt{1-\theta}}{\sqrt{1+\theta}}\right)$

$T_{\nu+1} =$ univariate Student's t CDF with $\nu+1$ degrees of freedom

PDFs that have an elliptical PDF and whose marginals follow a standard normal distribution.

Note in figure 4.3.1 that even though Gaussian copula and Student-*t* copula are radially symmetrical and have the same dependence parameter, their shape is different thanks to the tail dependence of the Student-*t* copula. Tail dependence will play a key role in reliability analysis, as shown in section 7.

4.3.2 Archimedean copulas

Before defining Archimedean copulas, the concept of generator function must be introduced. Following Nelsen [11] and Embrechts et al. [12], a *generator function* φ can be defined as a continuous strictly decreasing function that maps from $[0,1]$ to $[0,\infty]$, such that $\varphi(1) = 0$.

A generator function has a quasi-inverse $\varphi^{(-1)}$ that maps from $[0,\infty] \to [0,1]$, as follows:

$$\varphi^{(-1)}(t) = \begin{cases} \varphi^{-1}(t), & 0 \le t \le \varphi(0) \\ 0, & \varphi(0) \le t \le \infty. \end{cases}$$

Clearly, $\varphi^{(-1)}$ is continuous and decreasing on $[0,\infty]$ and it is also strictly decreasing on $[0,\varphi(0)]$. Thus, on the interval $[0,1]$,

$\varphi^{(-1)}(\varphi(t)) = t$. In this way:

$$\varphi\left(\varphi^{(-1)}(t)\right) = \begin{cases} t, & 0 \leq t \leq \varphi(0) \\ \varphi(0), & \varphi(0) \leq t \leq \infty. \end{cases}$$

A bivariate Archimedean copula is defined as:

$$\mathtt{C}(u,v) = \varphi^{(-1)}(\varphi(u) + \varphi(v))$$

where $\mathtt{C}$ is a copula if and only if φ is convex. Note that if $\varphi(0) = \infty$, then φ is a *strict generator*, and thus, $\varphi^{(-1)} = \varphi^{-1}$, which constitutes a *strict Archimedean copula* $\mathtt{C}(u,v) = \varphi^{-1}(\varphi(u) + \varphi(v))$.

Archimedean copulas constitute one of the most popular and flexible classes of copulas. The ease to construct and sample from these copulas and their wide variety of dependence structures, make this class of copulas a suitable option to model dependence in geotechnical engineering.

Although Archimedean copulas overcome some limitations of elliptical copulas, they have some disadvantages. Unlike elliptical copulas, Archimedean copulas are not derived from the *Sklar's theorem*; however, they do have closed-form expressions and they can model both radial and non-radial symmetry. Despite Archimedean copulas are especially suitable for modeling dependence of two random variables, their extension to more dimensions is not as straightforward as in the case of elliptical copulas. Actually, for extending this class of copulas to more than two dimensions some requirements must be fulfilled, and even so, there are some limitations over these kinds of procedures.

Nonetheless, the complexity of the final Archimedean copula increases as the dimension grows. Additionally, these extensions are too restrictive since the same dependence structure is applied to all random variables, which may be different from reality. In this regard, some modern techniques such as vine copula theory (see section 2.7) overcome these types of disadvantages and are more suitable for modeling high-dimensions dependence structures when using Archimedean copulas. Readers interested in a succinct review of the extension of Archimedean copulas to more than two dimensions are referred to Joe [32] and the references therein.

Many generator functions can be found in the literature, and hence many Archimedean copulas can be constructed. Among all Archimedean copulas, there are four widely used in geotechnical engineering for modeling dependence, namely Frank, Gumbel, Clayton, and Nelsen No 16 copulas. Among other similarities, the aforementioned four copulas have in common the capacity of modeling a wide spectrum of dependence, which is especially desirable in a copula to be used in geotechnics.

In particular, Frank copula is a comprehensive copula, Gumbel and Clayton copulas can model a complete spectrum of positive dependence and No 16 copula can model a complete spectrum of negative dependence and even a weak positive dependence. Additionally, there are some variants of these copulas, which extend their applicability to other structures of dependence; for example, Clayton copula can be modified to model both positive and negative dependence; also, survival copulas can be used or even a rotated version of the original copulas can be applied. These variants are not covered here, but readers interested in this topic are encouraged to refer to Nelsen [11].

As the generator function is the keystone of Archimedean Copulas, it can be used to estimate some of the copula properties. For example, copula parameter of dependence can be estimated from Kendall's tau using the respective copula generator function, and the same for the coefficients of tail dependence. Thus, complex equations such as the ones of tail dependence (equations (6) and (7)), or the one that relates Kendall's tau and the copula parameter of dependence (see section 4.4) can be simplified.

When the first derivative of φ^{-1} is equal to minus infinity, i.e. $\varphi^{-1'}(0) = -\infty$, coefficients of tail dependence for Archimedean copulas can be defined as:

$$\lambda_u = 2 - 2\lim_{s\to 0}\left[\frac{\varphi^{-1'}(2s)}{\varphi^{-1'}(s)}\right] \quad \text{and}$$

$$\lambda_l = 2\lim_{s\to\infty}\left[\frac{\varphi^{-1'}(2s)}{\varphi^{-1'}(s)}\right]$$

Conversely, if $\varphi^{-1'}(0)$ is finite, then both coefficients of tail dependence will be equal to zero. Proofs of these formulas are found in Genest and MacKay [34], Joe [32] and Embrechts et al. [12].

Table 3 presents some useful information about the most common Archimedean copulas in geotechnical engineering, i.e. Frank, Gumbel, Clayton, and Nelsen No. 16 copulas. Furthermore, there are several algorithms reported in the literature for sampling from different Archimedean copulas. A good exposure of these algorithms can be found in Nelsen [11], Embrechts et al. [12] or Phoon and Ching [35].

Coefficients of tail dependence of the studied Archimedean copulas are summarized in table 4.3.2.

For comparative purposes, contour plots of the bivariate PDFs generated through these four copulas are presented in figure 4.3.2. A coefficient of Kendall's tau $\tau = 0.3$ is used to calibrate these copulas, which is the same used for calibrating Gaussian and Student-t copulas of figure 4.3.1. Furthermore, the same marginal distributions, i.e. standard normal, are also used for constructing figure 4.3.2.

Note that the contour plots of the different copulas in figure 4.3.2 differ considerably, and these in turn differ from those of figure 4.3.1. These differences will be reflected in the probabilities of failure for any performance function, as we will see in section 7.

4.3.3 Plackett copula

Plackett copula is another widely applied copula in geotechnical engineering. This copula belongs to a family of its own, and it is a particular example of copulas constructed through algebraic methods. As the theory of the algebraic methods for constructing copulas is out of the scope of this document, readers are encouraged to refer to Nelsen [11], where a good introduction to this topic is presented.

Table 3 Generator function φ, copula function $\mathtt{C}$, and copula density function $\mathtt{c}$ in a bivariate case for some popular Archimedean copulas used in geotechnics.

Copula	$\mathtt{C}(u,v;\,\theta)$	$\mathtt{c}(u,v;\,\theta)$	$\varphi(t)$
Frank	$-\frac{1}{\theta}\ln\left[1+\frac{\left(e^{-u\theta}-1\right)\left(e^{-v\theta}-1\right)}{e^{-\theta}-1}\right]$	$\frac{-\theta\left(e^{-\theta}-1\right)e^{-\theta(u+v)}}{[(e^{-\theta}-1)+(e^{-u\theta}-1)(e^{-v\theta})-1]^2}$	$-\ln\left[\frac{e^{-t\theta}-1}{e^{-\theta}-1}\right]$
Gumbel	$\exp\left[-\left((-\ln u)^{\theta}+(-\ln v)^{\theta}\right)^{1/\theta}\right]$	$\frac{(-\ln u)^{\theta-1}(-\ln v)^{\theta-1}\left[s^{\frac{2-2\theta}{\theta}}-(1-\theta)s^{\frac{1-2\theta}{\theta}}\right]}{uve^{s^{1/\theta}}}$; $s=(-\ln u)^{\theta}+(-\ln v)^{\theta}$	$(-\ln t)^{\theta}$
Clayton	$\left[u^{-\theta}+v^{-\theta}-1\right]^{-1/\theta}$	$\frac{(1+\theta)(uv)^{-(\theta+1)}}{\left(u^{-\theta}+v^{-\theta}-1\right)^{\frac{1+2\theta}{\theta}}}$	$\frac{1}{\theta}\left(t^{-\theta}-1\right)$
No 16	$\frac{1}{2}\left(s+\sqrt{s^2+4\theta}\right)$; $s=u+v-1-\theta\left(\frac{1}{u}+\frac{1}{v}-1\right)$	$\frac{1}{2}\left(1+\frac{\theta}{u^2}\right)\left(1+\frac{\theta}{v^2}\right)z^{-1/2}\ldots$ $\ldots\times\left\{-\frac{1}{z}\left[u+v-1-\theta\left(\frac{1}{u}+\frac{1}{v}-1\right)\right]^2-1\right\}$; $z=\left[u+v-1-\theta\left(\frac{1}{u}+\frac{1}{v}-1\right)\right]^2+4\theta$	$\left(\frac{\theta}{t}+1\right)(1-t)$

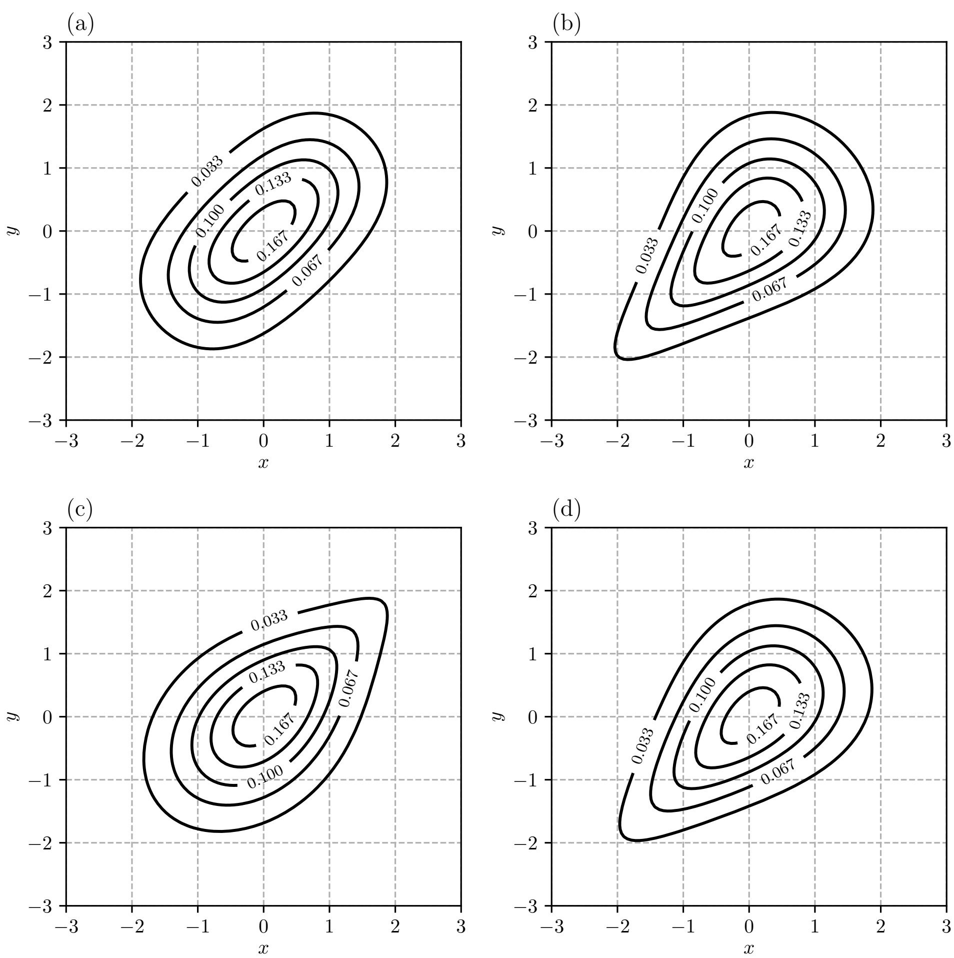


Figure 5 Contour plots of four bivariate Archimedean copulas with standard normal marginal distributions. (a) Frank copula ($\tau=0.3$), (b) Nelsen No 16 copula ($\tau=0.3$), (c) Gumbel copula ($\tau=0.3$), and (d) Clayton copula ($\tau=0.3$).

Table 4 Coefficients of tail dependence for some popular Archimedean copulas used in geotechnics.

Copula	λ_l	λ_u
Frank	0	0
Gumbel	0	$2-2^{1/\theta}$
Clayton	$2^{-1/\theta}$	0
No 16	0.5	0

A bivariate Plackett copula has the advantage of being a comprehensive copula, which is quite useful for modeling the dependence of geotechnical random variables. Additionally, among other properties, Plackett copula does not have tail dependence, either positive or negative, and is radially-symmetric, just like the Gaussian or Student-*t* copulas.

However, the extension of Plackett copula to more than two dimensions is not as simple as in the case of elliptical copulas, and this fact constitutes one of its disadvantages. Although some researches, such as Zhang and Singh [36], have presented an extension of Plackett copula to three-dimensions, its complexity increases significantly, like its limitations. In this regard, it is worth mentioning again that vine copula theory is especially helpful for modeling dependence in high-dimensions, and we recommend its use whenever necessary.

Additionally, there is no closed-form expression in the relation between rank coefficients and the Plackett copula parameter of dependence. Therefore, when adjusting a Plackett copula to a dataset, the parameter of the copula must be approximated through numerical methods. More of this adjustment will be explained in section 4.4.

Table 5 presents the equations of the bivariate Plackett copula and its associated PDF.

Table 5 Bivariate copula function $\mathtt{C}$ and copula density function $\mathtt{c}$ for the Plackett copula.

Copula	$\mathtt{C}(u,v;\,\theta)$	$\mathtt{c}(u,v;\,\theta)$
Plackett	$\frac{s-\sqrt{s^2-4uv\theta(\theta-1)}}{2(\theta-1)}$ $s=1+(\theta-1)(u+v)$	$\frac{\theta[1+(\theta-1)(u+v-2uv)]}{\left\{[1+(\theta-1)(u+v)]^2-4uv\theta(\theta-1)\right\}^{3/2}}$

Coefficients of tail dependence of the Plackett copula are summarized in table 6.

Table 6 Coefficients of tail dependence for the Plackett copula

Copula	λ_l	λ_u
Plackett	0	0

Algorithms for sampling from a bivariate Plackett copula can be found in Nelsen [11] or Li et al. [37]. Additionally, as it has been done with the other copulas of the different families, the contour plot of the bivariate PDF generated from of a Plackett copula are presented. For comparative purposes, figure 6 is constructed using a Kendall's tau coefficient equal to 0.3 for two random variables distributed as standard normal probability functions.

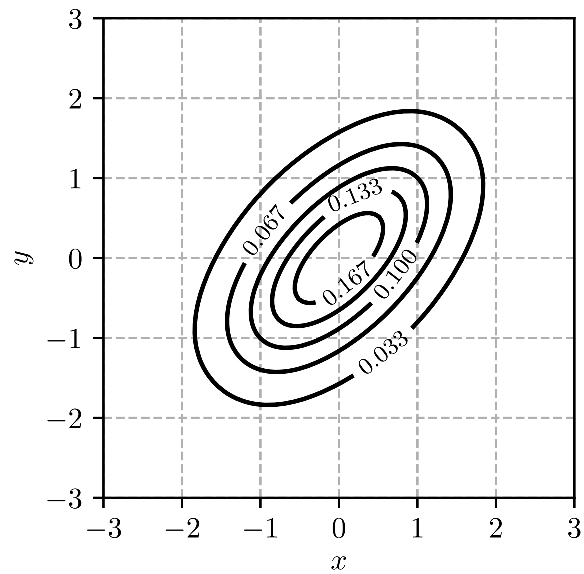


Figure 6 Contour plots of a bivariate Plackett copula for $\tau = 0.3$ with standard normal marginal distributions.

4.4 Estimation of the parameters of a copula

A copula might be specified in terms of one or several parameters that are gathered in a vector θ. In practice, θ is adjusted to a dataset using one of many available methods. In the following, an introduction to these methods is given, after McNeil et al. [5], Genest and Favre [38], and Durrleman et al. [39]. In particular, the method of moments and the maximum likelihood (with two of its variants: the IFM and the CML) will be explained.

4.4.1 Method of moments based on rank correlations

As stated in section 4.2, rank correlation statistics, like Kendall's tau or Spearman's rho, are the natural measures of dependence in copula theory. Even, it is possible to relate both Kendall's tau or Spearman's rho to the parameter of dependence of a copula.

When a copula $\mathtt{C}$ is absolutely continuous, its Kendall's tau correlation coefficient τ is defined using the Stieltjes integral [40, 11]:

$$\tau(\mathtt{C}) = 4\int_0^1\int_0^1 \mathtt{C}(u,v)\,\mathrm{d}\mathtt{C}(u,v) - 1. \tag{15}$$

In the same way, Spearman's rho correlation coefficient ρ_s can be defined for an absolutely continuous copula $\mathtt{C}$, as [40]:

$$\rho_s(\mathtt{C}) = 12\int_0^1\int_0^1 (\mathtt{C}(u,v) - uv)\,\mathrm{d}u\,\mathrm{d}v. \tag{16}$$

Observe that a reverse procedure is also possible in equations (15) and (16). That is to say, from a given Kendall's tau or Spearman's rho, the dependence parameter of a copula could be estimated. However, it is not usual to have the exact value of the rank measure but rather an approximation so, in this case, the dependence parameter defined through equation (15) or (16) is also an approximation.

In some cases, there exist simplifications of equations (15) and (16). Nonetheless, unfortunately in some other cases, these simplifications are not possible, and even equations (15) and (16) may not have an analytical solution.

In the case of Archimedean copulas, equation (15) can be simplified as a one-dimensional integral in terms of the generator function $\varphi_{\mathtt{C}}$ and its derivative, as [34]:

$$\tau(\mathtt{C}) = 1 + 4\int_0^1 \frac{\varphi_{\mathtt{C}}(t)}{\varphi'_{\mathtt{C}}(t)}\,\mathrm{d}t. \tag{17}$$

Furthermore, for some Archimedean copulas, such as Gumbel and Clayton copulas, equation (17) can be further simplified, obtaining a quite simple algebraic expression (see table 7). Nonetheless, not all Archimedean copulas achieve this level of simplification, such as in the case of Ali-Mikhail-Haq copula [see, e.g., 11], so numerical methods have to be employed over equation (17) or equation (15).

Regarding elliptical copulas, Pearson's rho correlation coefficient is their main descriptor of dependence, i.e. their copula parameter of dependence. However, as stated in section 4.2, this coefficient has many disadvantages from the copula's theory viewpoint. To overcome these disadvantages, relationships between Pearson's rho and rank measures of dependence can be used, such as the ones of equations (15) or (16). These estimates lead to the already presented equation (14), that relates Kendall's tau and the parameter of dependence of elliptical copulas. Additionally, there also exists a direct estimation of the elliptical copula parameter of dependence from Spearman's rho coefficient, which is given by [5]:

$$(\rho_s)_{ij} = \frac{6}{\pi}\arcsin\left(\frac{\rho_{ij}}{2}\right) \tag{18}$$

where ρ_{ij} is the ij-th element of the correlation matrix that contains the copula parameters of dependence. However, it is worth saying that equation (18) results in good approximations for Gaussian copulas, but this is not the case of other elliptical copulas such as the Student-t [41]. In this sense, it is preferable to employ the relationship $\tau - \rho$, given by the equation (14), since it is more robust and applicable to all elliptical copulas [5].

As stated before, not all copulas have a closed-form solution of equations (15) or (16), such as the Plackett copula. In these cases, numerical methods have to be employed to approximate the copula parameter of dependence by employing rank measures.

Table 7 summarizes the simplified equations that relate Kendall's tau τ and the copula parameter θ for the most commonly used copulas in geotechnics. When the relationship is not specified, it means that equation (15) must be solved using numerical methods. Additionally, table 7 also presents information on the range of τ values that the studied copulas in section 4.3 can model, as well as the range of values that their respective dependence parameter can adopt.

Table 7 Simplified relationships between Kendall's tau τ and the copula parameter θ for some widely employed copulas in geotechnics.

Copula	Range of τ	Relationship $\tau-\theta$	Range of θ
Gaussian	$[-1,1]$	$\tau=\frac{2}{\pi}\sin^{-1}(\theta)$	$[-1,1]$
Student-t	$[-1,1]$	$\tau=\frac{2}{\pi}\sin^{-1}(\theta)$	$[-1,1]$
Plackett	$[-1,1]$	–	$(0,\infty)\backslash\{1\}$
Frank*	$[-1,1]\backslash\{0\}$	$\tau=1-\frac{4}{\delta}+4\frac{D_1(\delta)}{\delta}$	$(-\infty,\infty)\backslash\{0\}$
Gumbel	$[0,1]$	$\tau=1-\frac{1}{\theta}$	$[1,\infty)$
Clayton	$[-1,1]\backslash\{0\}$	$\tau=\frac{\theta}{2+\theta}$	$[-1,\infty)\backslash\{0\}$
No. 16	$[-1,1/3)$	–	$[0,\infty)$

*Debye function $D_1(\delta)=\int_0^\delta \frac{x/\delta}{e^x-1}\,\mathrm{d}x$

4.4.2 Maximum Likelihood Estimator

A well-known alternative to the method of moments is the method of maximum likelihood estimation (MLE), which may be more robust and suitable than the method of moments when the information is scarce [20, 38]. This method allows to estimate parameters of a probability distribution by maximizing its likelihood function based on the available data.

In the framework of copula theory, the MLE method maximizes concurrently both marginal distributions parameters and copula parameters. In other words, by a single optimization, the most probable points in the parameter space for both structures are obtained.

Following Durrleman et al. [39], consider a continuous case for both the copula function $\mathtt{C}$ and all marginal distributions F_{X_i}, where $i = 1,2,\dots,d$, and d stands for the number of random variables. Now, assuming that the copula density $\mathtt{c}$ exists, and hence the joint PDF is given by equation (3), the expression of the log-likelihood can be written as [42]:

$$l(\theta) = \sum_{t=1}^{N} \ln \mathtt{c}\left(F_1\left(x_1^t;\,\theta_1\right), F_2\left(x_2^t;\,\theta_2\right)\dots F_i\left(x_d^t;\,\theta_d\right);\,\alpha\right) + \sum_{t=1}^{n}\sum_{i=1}^{d} \ln f_i\left(x_i^t;\,\theta_i\right) \tag{19}$$

where x_i^t is the d-dimensional sample of size N, θ_i and α are the vectors of parameters of the marginal distributions and the copula function, respectively, and $\theta = (\theta_1, \theta_2, \dots, \theta_d;\,\alpha)$.

By maximizing equation (19), the parameter vector θ is obtained, which corresponds to the maximum likelihood estimative. That is to say, θ is the parameter vector, from both copulas and marginal distributions, that maximizes the likelihood function of the model.

It is worth mentioning that the availability of the data is mandatory to carry out the MLE method since otherwise there is no way to maximize equation (19). By contrast, the method of moments based on rank correlations can be carried out by using the data or directly by a known Kendall's tau / Spearman's rho. Furthermore, note that both MLE method and moments method are strongly influenced by the amount and quality of data, since the greater amount/quality of data the more accurate estimates are obtained.

In some cases, the MLE method may seem to be less attractive than the method of moments, since it may involve a higher computational calculation cost over the copula density function and marginal distributions at the same time. However, this should not be a problem considering current computational capabilities.

On the other hand, the applicability of the MLE method is greater than the one of the method of moments, since the MLE method is especially useful when the copula has more than one fitting-parameter [38]. In this way, the degrees of freedom of a Student-t copula can be adjusted using the MLE, which is not possible using the method of moments. Thus, the MLE is a more comprehensive strategy than the method of moments, since the latter only adjusts the copula parameter of dependence and needs the help of other methods to adjust the rest of the parameters, in case they exist.

4.4.3 The inference functions from margins and the canonical maximum likelihood methods

Although the MLE is a quite robust and widely applicable method, it may become unfeasible in very high dimensions. The MLE method estimates both parameters of the marginal distributions and parameters of the copula function at the same time, and therefore it may be too computationally expensive. In this way, following the nature of the maximum likelihood methodology, it is necessary to explore other methodologies that are computationally more efficient but at the same time more robust and widely applicable.

The *inference functions for margins method* (IFM), appears as an alternative for estimating both copula function parameters and marginal distributions parameters [42]. The IFM method differs from the MLE method in the sense that the optimization is not done at the same time for the copula function and the marginal distributions. Instead, this procedure splits the MLE method into two stages. The first one fits the parameters of the univariate marginals distributions $\hat{\theta}_i$ using

$$\hat{\theta}_i = \arg\max_{\theta} \sum_{t=1}^{n} \ln f_i\left(x_i^t;\, \theta_i\right)$$

and in the second one, the copula vector of parameters α is computed based on $\hat{\theta}_i$, as follows

$$\hat{\alpha} = \arg\max_{\theta} \sum_{t=1}^{n} \ln \mathrm{c}\Big(F_1\left(x_1^t;\, \hat{\theta}_1\right), F_2\left(x_2^t;\, \hat{\theta}_2\right) \\ \dots F_d\left(x_d^t;\, \hat{\theta}_d\right);\, \alpha\Big)$$

As we can see, the IFM method follows the same nature of the MLE method, so they both share the same advantages over the method of moments. However, the IFM method is less computationally expensive than the MLE method, so its use in practice is sometimes preferred.

Additionally, following the idea of maximum likelihood estimation, there is a third technique known as the *canonical maximum likelihood* (CML) method. In this method, copula parameters are estimated without specifying marginal distributions, and for this purpose, empirical distributions are used to transform the original data $\left\{\left(x_1^t, \dots x_d^t\right)\right\}_{t=1}^{n}$ into values in the uniform space $\left\{\left(u_1^t, \dots u_d^t\right)\right\}_{t=1}^{n}$. In this sense, we can rewrite equation (19) as follows:

$$\hat{\alpha} = \arg\max_{\alpha} \sum_{t=1}^{n} \ln \mathrm{c}\left(\hat{u}_1^t, \hat{u}_2^t \dots \hat{u}_d^t;\, \alpha\right)$$

where $\hat{u}_n^t$ is obtained through the empirical marginal distribution as follows:

$$\hat{u}_n^t = \frac{R_t}{n+1} \tag{20}$$

in which R_t is the rank of the sample t and n is the total number of samples.

As a remark, these three maximum likelihood based methods (i.e. MLE, IFM and CML) result in quantities quite similar but with some degree of discrepancy. Especially, differences are more pronounced between the CML method and the other two methods, since the MLE and IFM methods do take into account a parametric function for the marginal distributions and the CML method does not. According to Durrleman et al. [39], if there exists a considerable difference between the results of the CML method and the ones of the MLE-IFM methods, these discrepancies may be indicative that marginal distributions are not well defined. In this case, the criteria for adjusting/selecting marginal distributions must be checked.

4.5 Tests of goodness-of-fit for copulas

Let us consider a sample $\left\{x_i \in \mathbb{R}^d : i = 1, 2, \dots, N\right\}$, and a set of copulas $\mathscr{C}$ fitted to that sample through any method exposed in section 4.4. The question that arises is which copula in $\mathscr{C}$ provides the best representation of dependence for the sample? In this regard, it is necessary to apply some criteria to evaluate which copula among the candidate copulas $\mathscr{C}$ is the most suitable.

In the case of marginal distributions, there are several popular goodness-of-fit tests, such as the Chi-Square (χ^2), the Kolmogorov-Smirnov (K-S), or the Anderson-Darling (A-D) tests [see, e.g., 20, for details]. However, in the case of copulas, there are no widely used and accepted methods such as in the selection of the marginal distributions. Attention has been commonly focused on how to fit copulas to data (as seen in section 4.4) and not on how to choose the most appropriate copula for a data set. For this reason, the selection of the best fit copula, especially when the information is scarce, is an open question in the state-of-the-art [43].

Following Genest et al. [43], methods that seek to select the most appropriate copula for representing the dependence of a d-dimensional sample can be divided into four categories:

- Special procedures for a certain families of copulas, like in the case of Archimedean or Gaussian copulas [see, e.g., 44, 45, 46].

- Procedures applicable to any type of copula, but whose implementation require an arbitrary parameter, kernel-weight functions or ad hoc categorization of the data in order to apply an analogous of the chi-square test.
- Procedures applicable to any type of copula, in which it is not required the adoption of any parameter or special strategy. These methods are called *blanket tests*, and their application is based on some transformations such as the Kendall's transform, the Rosenblatt's transform or the empirical copula [see 43].
- Standard goodness-of-fit tests, whose implementation is aimed at choosing the most appropriate copula in a set of candidate copulas $\mathscr{C}$, but not to test whether the selected model is suitable in the light of some measure such as the P-value [see, e.g., 20, 28, for its definition].

During the review of this state-of-the-art, we found that the most used tests for selecting the best copula in geotechnical engineering, and almost the only ones employed, belong to this last category. Thus, standard goodness-of-fit tests, specifically the Akaike-Information-Criterion (AIC) and the Bayesian-Information-Criterion (BIC), are widely used for selecting the better copula in $\mathscr{C}$ when representing the dependence of geotechnical variables. However, these two tests do not validate the choice of the copula.

To clarify this concept, take into account that standard goodness-of-fit-tests seek to define the copula, within a set of candidate copulas $\mathscr{C}$, that best represents the available information. However, if all copulas in $\mathscr{C}$ are poor representatives of the sample dependence, the least poor representation of dependence (i.e., copula) will be selected. Therefore, all copulas in $\mathscr{C}$ must be responsibly proposed, taking into account some characteristics such as the range of dependence values that the copula can handle, tail dependence, symmetry, etc. For example, if two random variables have a negative correlation, $\mathscr{C}$ must contain copulas that can model negative dependence, and not copulas that are only restricted to positive dependence.

Other methods worth delving are the blanket tests, since these validate the suitability of a copula through some criteria such as the P-value. An excellent exposure of these methods is given by Genest and Favre [38] and Genest et al. [43], so readers are encouraged to consult these references. In this document we will introduce two blanket tests: the Cramér-von-Mises and Kolmogorov-Smirnov tests.

Next, section 4.5.1 is devoted to the Cramér-von-Mises and Kolmogorov-Smirnov goodness-of-fit tests, while section 4.5.2 deals with the Akaike Information Criterion and the Bayesian Information Criterion.

4.5.1 Goodness-of-fit test based on the empirical copula

Starting from the concept of empirical copula already introduced in section 2.6, it is possible to apply analogous methodologies to the univariate Cramér-von-Mises and Kolmogorov-Smirnov goodness-of-fit tests in order to evaluate the suitability of a copula function when modeling the dependence of a data sample. For this purpose, it is necessary to compare the distance between the empirical copula $\mathtt{C}_N$ and the parametric copula $\mathtt{C}_\theta$ under the null hypothesis H_0, i.e., the evaluated copula model cannot be rejected.

Following Genest et al. [43], Cramér-von-Mises test statistic S_N, employing an empirical copula, is defined as:

$$S_N = \int_{[0,1]^d} \left(\sqrt{N}(\mathtt{C}_N(u) - \mathtt{C}_\theta(u))\right)^2 \mathrm{d}\mathtt{C}_N(u) \tag{21}$$

Analogously, Kolmogorov-Smirnov test statistic (T_N) can be expressed as:

$$T_N = \sup_{u\in[0,1]^d} \left|\sqrt{N}(\mathtt{C}_N(u) - \mathtt{C}_\theta(u))\right|. \tag{22}$$

In the above equations, N is the sample size and $\mathtt{C}_N(u)$ is the empirical copula constructed through equation (8).

When $\mathtt{C}_\theta$ has an analytical expression, then S_N and T_N can be directly computed using equations (21) and (22), respectively; otherwise, a Monte Carlo simulation with $m \geq N$ trials would have to be applied, as presented in Genest et al. [43].

4.5.2 Akaike and Bayesian information criterion

The Akaike Information Criterion [47], and the Bayesian Information Criterion [48], hereinafter referred to as AIC and BIC correspondingly, are criteria of model selection within a finite set of candidate models. That is to say, both AIC and BIC evaluate the quality of a model and compare it with the quality of the other candidate models that represent the data set. In practice, these two criteria are closely related, although their origin and nature is different [49, 50].

Akaike information criterion is defined as:

$$AIC = 2k - 2l \tag{23}$$

meanwhile Bayesian information criterion is given by:

$$BIC = k \ln N - 2\ln l; \, , \tag{24}$$

here k represents number of model/copula fitting parameters ($k = 1$ for a parametric copula where θ is a one-dimensional vector), l stands for the value of the likelihood function of the copula density function adjusted with θ; N is the sample size, equivalent to the number of observations or measures.

In both cases, the copula $\mathtt{C}$ with the smallest AIC/BIC in the set $\mathscr{C}$ is the preferred copula. Equations (23) and (24) can be divided into two parts, one that corresponds to the adjustment of the model to the available data, represented by the value L of the likelihood function, and the other corresponding to the number of parameters used by the model, represented by k. Thus, both tests have a preference for the models that best fit data, but in turn, penalize the complexity of the model given by the number of adjusted parameters. Therefore, both tests look for models that neither overfit nor underfit the true model.

Although these two tests are quite similar, and even the AIC can be derived from a Bayesian framework just like the BIC [49], there are several differences between them. For example, penalty for the number of parameters is greater in the case of the BIC test in comparison to the AIC test. Additionally, some authors argue that the use of the AIC or BIC tests must be evaluated depending of the target task [51, 49].

Particularly, the BIC test is argued as especially useful for selecting the "true model". Thus, if the true model is in the set of candidate models, the BIC test will select it with a probability of one when the sample size tends to infinity. Nonetheless, in geotechnical engineering practice, there is no true or perfect model since these are just approximations of reality, and as an aggravating factor, sample sizes are usually small. Therefore, under these circumstances, the AIC has shown a better performance than the BIC.

Nonetheless, as the formulas of AIC and BIC tests are easily calculable, it is a good practice to compute both and compare the selected best model within the candidate models set. Usually, the best model will be the same independently of the used test. A detailed comparison between these two methods is exposed by Burnham and Anderson [51, 49], Vrieze [50].

5 Some applications of copulas in geotechnical engineering

This section is devoted to present several uses of copula theory in geotechnical engineering, in concordance with the current state-of-the-art. Section 5.1 describes how they have been employed to model the dependence between shear strength parameters. Section 5.2 discusses the applicability of copulas for modeling settlement parameters. The uses of copulas in geotechnical earthquake engineering and soil spatial interpolation are presented in sections 5.3 and 5.4, respectively. Other uses of copulas in relation to geotechnics are presented in section 5.5. Since reliability analysis is one of the main applications of copulas in geotechnics, section 7 will be devoted to discuss this topic.

5.1 Shear strength parameters of the Mohr-Coulomb failure criterion

In geotechnical engineering, Mohr-Coulomb failure criterion is a popular mathematical model that describes the response of soil and rock materials to shear and normal stresses. This model is represented by a linear envelope that relates the shear strength of the material and the applied normal stress, as:

$$\tau = \sigma \tan\phi + c \tag{25}$$

where τ denotes the shear strength, σ is the normal stress, and c and ϕ are the shear strength parameters, cohesion and friction angle, respectively. Note that in equation (25), cohesion is the intercept of the envelope with the τ axis, and $\tan\phi$ is the slope of the envelope.

Both cohesion and inner friction angle are commonly obtained from the same laboratory test, either by a direct shear test or by a triaxial test [see 52, 53, for details of these two tests and their modalities]. For this reason, and since both parameters come from the same soil or rock sample, there exists an intrinsic relation between them that can be studied from a probabilistic perspective. Some researches have pointed out a negative dependence between cohesion and friction angle [see, e.g., 54, 55, 56], and others a positive one [see, e.g., 57, 58]. Whatever the case, this dependence exists and cannot be neglected, as stated by Lumb [57], Matsuo and Kuroda [59], Cherubini [60], Fenton and Griffiths [61] or Fellin and Oberguggenberger [62]. Nonetheless, some researches have worked with cohesion and friction angle as independent parameters [see, e.g., 63, 64, 65, 66, 67], although this is an unwarranted assumption.

A rudimentary approximation to the dependence between cohesion and inner friction angle is expressed in terms of Pearson's rho correlation coefficient [employed by, e.g., 68, 69, 70]. However, this approximation is not adequate considering the large dispersion between values of these two parameters and the disadvantages of Pearson's rho correlation coefficient that have already been exposed in section 4.2. Nonetheless, since Pearson's rho coefficient is quite simple to compute and is the basis of the popular multivariate normal distribution, it has been widely employed in geotechnical engineering [see 71, 72, for further details].

Given that the multivariate normal distribution is quite restrictive and in general has several disadvantages (see section 3), other alternatives had to be explored to model the dependence between soil shear strength parameters. The Nataf isoprobabilistic transformation was one of these alternatives, and in particular, this transformation was one of the most widely used approaches for modeling dependence before the development of copula theory. However, as exposed in section 4.2, the Nataf transformation is indeed a particular case of the Gaussian copula. Thus, in light of copula theory, soil shear strength parameters have been modeled employing only one dependence structure, the one given by the bivariate Gaussian copula.

To fill this gap, some researches have evaluated other alternatives to the Gaussian copula to model the dependence between shear strength parameters, such as the Plackett, Frank, or Nelsen No. 16 copulas [see, e.g., 73, 74, 75, 76, 77]. Note that the aforementioned copulas can handle a wide spectrum of negative and positive dependence. This property is needed since the dependence between soil shear strength parameters varies in a wide spectrum of negative values, and even sometimes positive values are reported in literature. Some popular copulas such as the Farlie-Gumbel-Morgenstern (FGM) or the Ali-Mihkail-Haq (AMH), which are restricted to a relatively weak dependence [11, 78], may not be suitable to model the dependence of soil and rock shear strength parameters.

Furthermore, it is worth mentioning that the copulas in the set of candidate copulas $\mathscr{C}$, for representing a dataset of soil shear strength parameters, must be varied and have different properties such as asymmetry, tail dependence, range of dependence values, etc. These differences allow to evaluate several possibilities in order to select the one that best fits the structure of dependence of the dataset.

As an illustrative example, in this section we are carrying out the exercise of defining a copula model for a soil shear strength dataset. This exercise will be executed based on the procedure studied in section 4 and it is in concordance with the current state-of-the-art on the construction of copulas for soil/rock shear strength parameters. Then, comparisons with other investigations of the same nature, and some remarks, that in the judge of the authors are worth highlighting, are also presented. It is worth mentioning that this exercise can be extended to other geotechnical parameters, so it is not limited to soil/rock shear strength parameters.

The soil shear strength dataset from the Xiaolangdi Hydropower Station was adopted as example in this section [readers can consult this dataset in 79]. This dataset is particularly useful since it has already been studied from a probabilistic perspective, and even from copula theory framework, in other works [see, e.g., 35, 80]. A comprehensive soil description and a probabilistic description of the data can be found in Yu [81] and Zhang et al. [79].

The Xiaolangdi Hydropower Station dataset contains measures of soil shear strength parameters from three loading conditions of triaxial tests, namely consolidated-drained test (CD), consolidated-undrained test (CU), and unconsolidated-undrained test (UU). The sample size of each one of these tests is equal to 63, 64, and 61, respectively.

As a preliminary step, it is necessary to study the type of dependence between cohesion and friction angle of each triaxial test modality. For this purpose, dependence can be isolated of the effect of marginal distributions by applying the empirical distribution over the original dataset.

Following this guideline, figure 7 contains three scatter plots in the uniform space for each loading condition of triaxial tests. This scatter plots are obtained after applying the transformation given by the empirical distribution, equation (20).

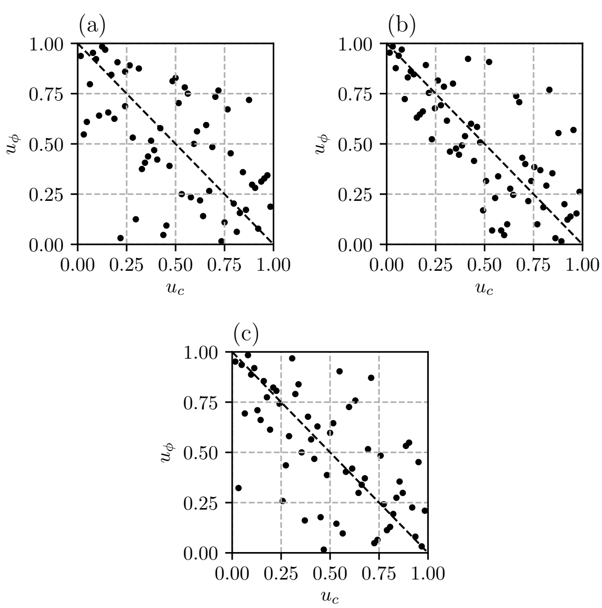


Figure 7 Scatter plots in the standard uniform space of c vs ϕ, from the dataset of the Xiaolangdi Hydropower Station after employing the transformation given by the empirical distribution. (a) CD dataset, (b) CU dataset, (c) UU dataset.

Scatter plots of figure 7 show that there exists a negative trend between cohesion and friction angle for all modalities of triaxial tests. Points of all the scatter plots of the figure 7 are located next to a $-45°$ slope line, which represents a perfect negative dependence.

The negative correlation between cohesion and friction angle for each triaxial test modality is validated through the respective Pearson's rho and Kendall's tau coefficients. The resulting Pearson's rho coefficients for the CD test, CU test, and UU test, are equal to -0.544, -0.702, and -0.623, respectively. The resulting Kendall's tau coefficients for the CD test, CU test, and UU test, are equal to -0.384, -0.544, and -0.447, respectively.

Considering this negative dependence, one might already have some candidate copulas in mind to model the dataset. For instance, the Gaussian, Plackett, Frank and Nelsen No. 16 would be good options. Furthermore, some other copulas that have not been extensively used, so far, to model shear strength parameters can be proposed. For instance, the Student-*t* Copula is a suitable option, although it has been only used in a few works of copulas in geotechnics [e.g., 76, 82]. Some researchers point out that the Student-*t* copula is more general than the Gaussian copula, and should be used instead of this last one when possible [see, e.g., 12].

It is worth mentioning that some other authors have even worked with copulas that are not suitable to model negative dependence in a first instance, such as the Gumbel or Clayton copulas [e.g. 82, 74]. For this purpose, they transform one of the random variables to force a positive dependence. This operation is also valid and allows to take advantage of the particular properties of interest of some copulas. Details of this transformation can be found in the above references or in Nelsen [11].

Now, the next step is to adjust the candidate copulas to the available data. Recalling section 4.4, there exist four practical methods to adjust the candidate copulas to the available information. The MLE method calibrates both copula function and marginal distributions at the same time, while the IFM, CML, and moment methods, calibrate marginal distributions and the copula function separately.

In the literature, these four fitting methods have been indiscriminately used for adjusting copulas to shear strength parameters. The MLE method was employed by Zhang et al. [83] to adjust marginal distributions and copulas at the same time. On the other hand, [84], [35], and [82] employed the IFM method to firstly adjust marginal distributions and posteriorly copulas. The CML was exposed in the work of [75]. Finally, the method of moments has been used by [74], [76], or [85].

It is possible to conclude that the most used methods for fitting copulas to datasets of shear strength parameters correspond to the IFM method and the method of moments. Note that unlike the IFM method, the method of moments only estimates the copula parameter of dependence, so other copula parameters and marginal distributions must be adjusted independently.

However, the importance of fitting copulas and marginal distributions at the same time, i.e., of using the MLE method to adjust copulas to data, was pointed out by Zhang et al. [83]. In this regard, Zhang et al. [83] expresses that copulas affect the fitting values of the marginal distributions, so for a robust estimation of the models, the MLE method must be used instead of the other methods.

We agree with [83] at considering the MLE method as the most robust method for fitting copulas to data; some other researches, such as Durrleman et al. [39], support this postulate. This consideration is based on the fact that data is obtained as samples and not as populations, so there will be discrepancies among the results of the four fitting methods. The scarcer the information, the greater the discrepancies between results of the methods; and vice versa, the more information, the smaller the differences among results of the methods. The MLE method, by fitting marginal distributions and copula at the same time, maximizes the likelihood function of the model and may lead to the best model estimate based on the available data; this is in comparison to other methods that adjust separately marginal distributions and copulas.

Nonetheless, when fitting copula models to data, Zhang et al. [83] allows negative values of soil cohesion, which is physically impossible. In this way, the work of Zhang et al. [83], and its conclusions, must be reviewed in light of this physical restriction. It is important to be aware of the physical restrictions of the random variables, since they has a direct impact on the final results. Thus, further studies should be carried out to quantify the discrepancies between the fit values of models, obtained through all the different methods used for this purpose.

As an example in this document, several marginal distributions will be adjusted to the data maximizing their likelihood function. These marginal functions are the truncated normal below zero, lognormal, gamma, beta, and Weibull distributions. Then, the suitability of each marginal distributions will be compared through goodness-of-fit-test.

There are several goodness-of-fit tests for univariate distributions [see e.g. 20, 28], and some of these have been widely applied to model the distributions of the shear strength parameters. In the copula framework for modeling shear strength parameters, the Kolgomorov-Smirnov test has been used by Huang et al. [84] or Xu et al. [74], and the Anderson-Darling was used by Wu [76, 77]. Even the AIC and BIC methods have been used as means of model selection of univariate distributions [e.g. 35, 75], although these do not constitute a formal goodness-of-fit test (see section 4.5.2).

In most cases, the Kolmogorov-Smirnov test is enough to assess the feasibility of a univariate probability model, and no other tests are needed. Thus, the K-S test is applied in this document to define the best-fit marginal distributions at a significance level of 0.05 ($\alpha = 0.05$). Results are summarized in table 5.1.

On the other hand, AIC/BIC values of all the marginal distributions were computed and used in the original work of Phoon and Ching [35, chap. 2] to define the best fitting model to data. As a comparative exercise, it is interesting to contrast best-fit marginal distributions obtained through these two methodologies, i.e. K-S test and AIC/BIC test. Table 5.1 presents the AIC/BIC values for all the parameters and marginal distributions.

In the first instance, it is evident that there are some minor discrepancies between some AIC/BIC values reported in this document compared with those reported by Phoon and Ching [35, chap. 2]. These discrepancies are mainly due to two reasons: the fitting method and the limits of the truncated functions. While the method of moments was used by Phoon and Ching [35, chap. 2], in this document the likelihood of each distribution was maximized. Therefore, it is expected that different fitting parameters were obtained, although close.

As an interesting remark, note that AIC/BIC values reported in this document are in all cases smaller than those reported by Phoon and Ching [35, chap. 2]. This fact indicates that maximizing the likelihood functions lead to better-fit models than those obtained by the method of moments used by Phoon and Ching [35, chap. 2].

Regarding best-fit marginal distributions, both methodologies agree on their definition in the majority of cases. Note that in this particular example, most suitable marginal distributions differ in the case of inner friction angle of the CD triaxial test, and in the case of cohesion of the UU triaxial test. Nonetheless, in these two cases, the best-fit distributions obtained through both methodologies have quite similar values of AIC/BIC or K-S, so their data representation is similar and their differences are minor, in practice.

Additionally, the present work includes two marginal distributions that were not evaluated in the study of Phoon and Ching [35, chap. 2], namely the gamma and beta probability distributions. Particularly, these two distributions are found suitable to model both shear strength parameters of the consolidated drained triaxial test. This find is a motivation to include different types of univariate distributions within the set of candidate distributions to represent cohesion and friction angle, or other geotechnical parameters.

After defining marginal distributions, the next step is proposing a set of copulas, adjust them to data, and select the most suitable one. For this purpose, the same copulas proposed by Phoon and Ching [35, chap. 2] are evaluated in this document, i.e. Gaussian, Plackett, Frank, and Nelsen No. 16 copulas. Additionally, Student-t copula is also evaluated since it is a more general copula than the Gaussian copula, and its use is recommended by some researches such as Embrechts et al. [12].

The parameter of the dependence of each copula is fitted through the method of moments, i.e., using equation (15) or its simplifications. Furthermore, the number of degrees of freedom of the Student-t copula is assumed to equal to 4.0 as recommended when it is not possible to estimate it [5].

Although in this case a maximum likelihood estimation can be carried out to compute the number of degrees of freedom of the Student-t copula, for illustrative purposes ν is assumed in order to represent tail dependence and compare the Student-t copula with the Gaussian copula.

Figure 5.1 presents scatter plots for the CD triaxial test of the Xiaolangdi Hydropower Station, in conjunction with 300 points simulated through the aforementioned 5 copulas, after their respective adjustment to the data. It is worth mentioning that for this simulation the best-fit marginal distributions defined in table 5.1 were employed.

Table 8 K-S values of the different marginals distributions. Here D_{max} stands for K-S test statistic.

Triaxial test	Parameter	Truncated normal	Lognormal	Gamma	Beta	Weibull	D_{max}
CD	c	0.072	0.127	0.102	**0.06**	0.067	0.171
	ϕ	0.109	**0.093**	0.098	0.146	0.151	0.171
CU	c	**0.056**	0.173	0.127	0.092	0.076	0.17
	ϕ	**0.082**	0.111	0.1	0.103	0.091	0.17
UU	c	**0.083**	0.151	0.127	0.124	0.088	0.174
	ϕ	0.082	0.065	**0.051**	0.091	0.072	0.174

Values in bold highlight the best-fit marginal distributions

Table 9 AIC/BIC values of different marginals distributions

Triaxial	Parameter	Truncated Normal		Lognormal		Gamma		Beta		Weibull	
		AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
CD	c	568.6	572.89	589.77	594.06	579.83	584.11	**567.93**	**572.22**	568.94	573.23
	ϕ	300.86	305.14	300.26	304.55	**300.15**	**304.44**	307.77	312.05	310	314.29
CU	c	**613.14**	**617.46**	645.97	650.28	627.48	631.79	617.58	621.9	617.01	621.33
	ϕ	**340.34**	**344.66**	342.11	346.43	341.03	345.35	341.75	346.07	343.85	348.16
UU	c	667.86	672.08	674.77	679	670.39	674.62	675.13	679.35	**667.49**	**671.71**
	ϕ	345.38	349.6	338.95	343.17	**338.77**	**342.99**	347.32	351.55	344.34	348.57

Values in bold highlight the best-fit marginal distributions

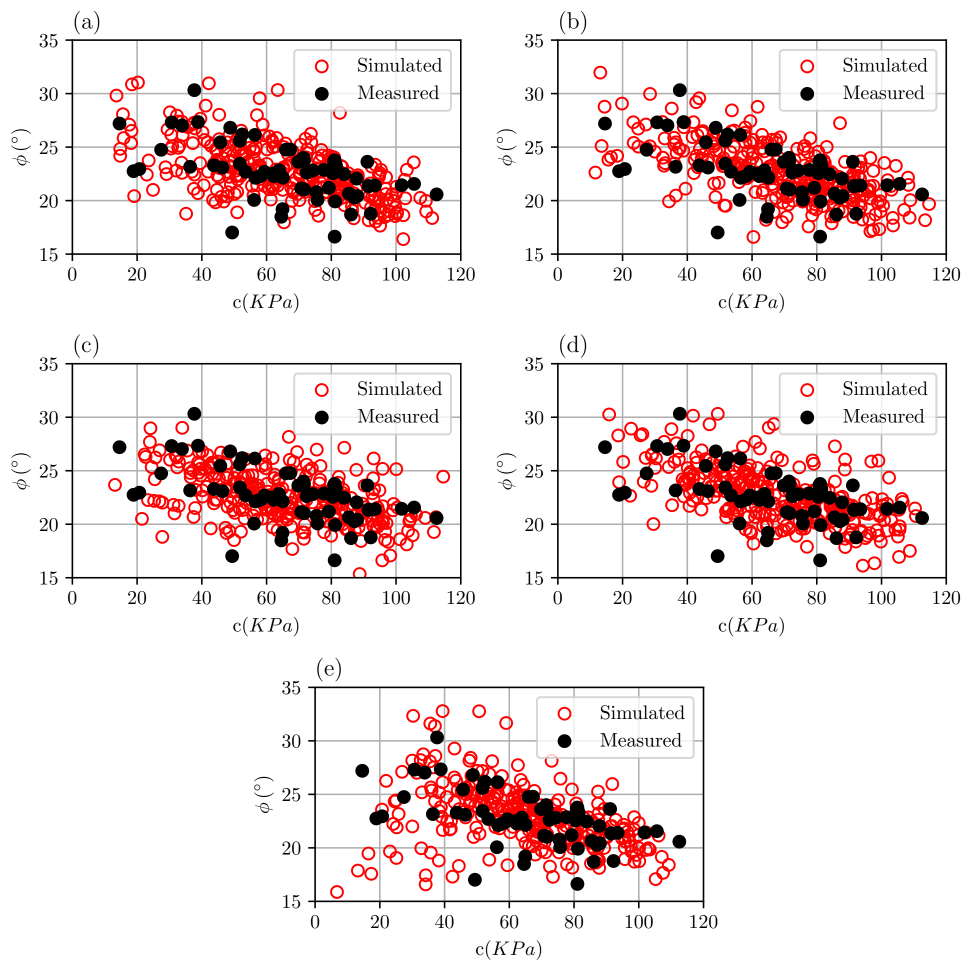


Figure 8 Simulation of CD data from different copulas. Cohesion modeled as a beta distribution and friction angle modeled as a lognormal distribution (a) Gaussian copula, (b) Student-*t* copula, (c) Plackett copula, (d) Frank copula, (e) Nelsen No 16 copula.

Table 10 AIC-BIC values for all the studied copulas

Triaxial Test	Gaussian		Student-t		Plackett		Frank		Nelsen No. 16	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
CD	-19.65	-17.51	-13.72	-9.44	-18.05	-15.91	**-20.22**	**-18.07**	-15.4	-13.26
CU	-43.62	-41.47	-39.72	-35.4	-43.12	-40.96	**-44.22**	**-42.07**	-32.46	-30.3
UU	-26.44	-24.33	-24.28	-20.06	**-27.46**	**-25.35**	-27.28	-25.17	-23.42	-21.31

Values in bold highlight the best-fit copulas

It seems that all copulas well-fit to data, but evidently their suitability must be validated through goodness-of-fit tests. For this purpose, the AIC or BIC values were computed for each copula and triaxial test. The results of these calculations are summarized in table 5.1.

From table 5.1, Frank copula is defined as the most suitable copula for the datasets of CD and CU triaxial tests, while the Plackett copula is defined as the most suitable copula for dataset of the UU triaxial test. This means that the Gaussian copula is not the best representation of dependence between soil shear strength parameters of the Xiaolangdi Hydropower Station.

Gaussian dependence structure is commonly adopted without validation in practice, but in the light of copula theory, this is just one of several dependence structures. Therefore, the suitability of the Gaussian copula for modeling dependence of a dataset must be evaluated and contrasted with other copulas. The same conclusion is found in the many studies of dependence between soil/rock shear strength parameters from the framework of copula theory.

Additionally, note that in this particular case Gaussian copula performs better than Student-t copula, as observed in the AIC/BIC values of table 5.1. These differences between the AIC/BIC values of the Gaussian and Student-t copula are due to the assumed degrees of freedom. In this way, shear strength parameters from the Xiaolangdi Hydropower Station are better represented with no tail dependence. This is corroborated with the best fit copulas since the coefficients of tail dependence are equal to zero for both Frank and Plackett copulas.

There exist other shear strength datasets that have been studied from a copula theory perspective. For instance, it is worth mentioning the datasets of the Ankang Hydropower station [see 85, 83], the dataset of the Jiangkou, Shuikou and Ertan hydropower stations [see 85], the dataset of the Shicuan hydropower station [see 84], a shear strength dataset of soils from Shanghai [see 86, 79], among others. Nonetheless, similar results and conclusions were obtained in all these works.

Furthermore, Wu [82] conducted a study, from copula theory perspective, in which several dataset of shear strength parameters were employed. For all these datasets, Wu [82] constructed copulas and found similar results and conclusions to those obtained in here. In general, there is no universal accepted copula for modeling shear strength parameters of soil and rocks. Each problem has its particular conditions, so copulas must be fitted to data and the best fit copula selected through the methods explained here in order to obtain the best possible representation of dependence.

Finally, it is worth mentioning that procedures described in this section are based on the current state-of-the-art. So, these procedures could be replicated to model any dataset of shear strength parameters or even any bivariate dependence in geotechnical engineering.

5.2 Settlement

In the framework of settlement analysis, copula theory has been mainly employed for the study of model uncertainty of different equations that are used for serviceability-limit-state designs. Most of these equations relate, for different types of foundations, the applied load and the consequent displacement. To this end, these relations are commonly expressed in terms of a normalized load Q and a normalized displacement η. Commonly, the normalized load is obtained through a load of reference, and the normalized displacement is obtained through some geometrical factor. After the normalization, the final scatter in the curves is significantly reduced, as shown by Dithinde et al. [87], Huffman et al. [88] or [89, 90]. In this regard, copulas have been used to model the final scatter of these curves, represented by the fitting parameters of their equations (i.e., model uncertainty).

Foundation load-displacement relationships can be evaluated through several equations reported in literature. In particular, two equations have been mostly employed, which are known as the hyperbolic equation and the power law equation. *Hyperbolic equation* can be expressed in normalized terms as:

$$Q = \frac{\eta}{a + b\eta} \tag{26}$$

while the *power law equation* can be expressed in normalized terms as:

$$Q = c\eta^d \tag{27}$$

where Q is the normalized load, η represents the normalized displacement, and a, b, c, and d, are model fitting parameters.

To normalize load-settlement curves several methodologies have been applied. In the case of the induced settlement, this is mostly normalized by a foundation geometry factor, commonly known as B. The parameter B can represent the diameter of the foundation, the width of the foundation, or even the equivalent footing diameter, i.e. the diameter that produces the same area as that of a square footing [see, e.g., 91]. The normalized settlement is hence computed as $\eta = s/B$, where s is the measured settlement and B the geometry normalization factor.

On the other hand, normalized load Q employs a reference capacity q_{ref} to normalize the applied load q. Reference capacity is not necessarily the true foundation capacity, but rather a known capacity, preferably obtainable through a deflection criterion [88].

For example, following Briaud [92], in-situ cone net tip resistance $q_{c,net}$, from the cone penetration test (CPT), was employed by [89, 90] as q_{ref} for normalizing applied loads on shallow footings on sand. On the other hand, following Phoon et al. [93, 94], the slope tangent capacity Q_{STC} was employed by Li et al. [95, 37] as q_{ref} for normalizing the applied load on static piles, or by Huffman et al. [88] over spread footings on clays. Details of the cone penetration tests (CPT) and in-situ cone net tip resistance $q_{c,net}$ are found in Robertson and Cabal [96] and Briaud [92]. More details about the slope tangent capacity Q_{STC} are found in Phoon et al. [93, 94].

It is worth mentioning that a normalized displacement is not always employed in equations (26) or (27) before modeling the fitting parameter of the curves through copulas. For example, only the applied load was normalized by Dithinde et al. [87] and Li et al. [95, 37] when employing equation (26) for representing load-displacement curves of static piles. In this case, η in equations (26) and (27) is replaced by the crude settlement measure s. Conversely, unlike normalized displacement, normalized load is always employed in equations (26) or (27) before modeling fitting parameters through copulas.

Copula theory has been employed for modeling the remaining scatter after normalizing the load-displacement curves dataset. In this way, dependence between parameters a and b from equation (26), or c and d from equation (27), has been studied under the copula theory framework.

For instance, Uzielli and Mayne [89] used a load-displacement database of 30 shallow footings in different sands to calibrate equation (27). After the calibration, they proposed a methodology to estimate the design load for different threshold settlements in shallow footings on sands, based on the CPT and for a given reliability level.

It was demonstrated by Uzielli and Mayne [89] that the parameters c and d of equation (27) are dependent on each other and independent of other variables such as B or q_{app}. They employed a lognormal distribution to model both c and d parameters, and different copulas were evaluated to represent their dependence. Frank copula was found to be the best-fit copula for c and d, and additionally, other variables such as foundation width B, q_{net} or q_{app} were also adopted as random variables but without dependence for the analysis. Thus, a probabilistic approach was presented by Uzielli and Mayne [89] to design vertically loaded shallow footings on sand based on the CPTu test, and taking into account model uncertainty of equation (27) with the support of copula theory.

Later, Uzielli and Mayne [90] delve into the importance of taking into account the dependence between fitting parameters of the power law model (equation (27)). The same load-displacement database of Uzielli and Mayne [89] was employed by Uzielli and Mayne [90] for calibrating equation (27). They evaluated three procedures for modeling load-displacement curves of shallow footings on sand, namely the fully probabilistic MP, the parametric uncertainties PO, and the deterministic DT approaches. MP approach consists of a fully probabilistic method, where B, q_{app}, and q_{net} parameters are modeled as independent random variables; additionally, both c and d parameters from equation (27) are modeled as random variables through a Frank copula and beta marginal distributions. The PO method comprises in a probabilistic method where B, q_{app} and q_{net} are independent random variables, but c and d in equation (27) are taken as deterministic values. Finally, DT approach involves a fully deterministic method where B, q_{app} and q_{net}, just like the parameters c and d, are taken as deterministic values.

In this way, Uzielli and Mayne [90] highlights the importance of probabilistic analysis when modeling load-displacement curves; in particularly, the importance of the dependence between c and d parameters of equation (27) is highlighted. They pointed out that the MP approach showed to be the most appropriate one, and neglecting the inherent uncertainty and dependence will result in inaccuracies. Settlement is overestimated in low loading levels and underestimated in high loading levels when the MP methodology is not employed.

Similar works were conducted by Huffman and Stuedlein [97] and Huffman et al. [88]. The former studied, from a reliability-based point of view, the limit state design of spread footings on aggregate pier reinforced clay. The latter, studied, again from a reliability-based point of view, the immediate settlement of spread footings on clays. These two works used copulas to model the dependence between two power-law parameters, i.e. c and d from equation (27), in order to incorporate uncertainties into the methodologies of study. We do not delve into these works, since the use of copulas is quite similar to that employed by Uzielli and Mayne [89, 90]. Readers are referred to Uzielli and Mayne [89], Uzielli and Mayne [90], Huffman and Stuedlein [97] and Huffman et al. [88] for further information.

On the other hand, several investigations that employ copula theory have been applied to the pile load-displacement database collected, organized, and categorized by Dithinde et al. [87]. The data was collected from various piling companies that operate in the Southern Africa countries, and it was grouped into four subsets of pile load tests, as follows: driven piles in non-cohesive soils (D-NC), bored piles in non-cohesive soils (B-NC), driven piles in cohesive soils (D-C), and bored piles in cohesive soils (B-C). The sample size of each category of pile load test is equal to 28, 30, 59 and 53, respectively. In addition to the load-displacement data, the properties of the soil and the geometric characteristics of the foundations are reported.

Dithinde et al. [87] demonstrated that the scatter in the original load-displacement curves can be substantially reduced when the applied load is normalized by the interpreted capacity of the pile [see 94]. This procedure leads to normalized curves that can be correctly fitted through the hyperbolic model. However, it is worth noting that they do not normalize settlement measures, so equation (26) is rewritten as

$$Q = \frac{s}{a+bs} \tag{28}$$

where Q is the normalized load, s the pile top displacement, and a and b are curve fitting parameters.

Curve fitting parameters from equation (28), a and b can be interpreted as the reciprocals of the initial slope and asymptotic value of the normalized curves. The procedure to obtain the interpreted capacity Q_i is detailed in the work of Dithinde et al.

[87], where a conjugation of Chin's extrapolation procedure [98] and Davisson's failure criterion [99] was employed.

After normalizing the applied load of the datasets, there is a remaining model uncertainty that can be represented modeling both curve fitting parameters. Particularly, a and b parameters of equation (28) are found to be mutually dependent and independent of other variables.

Figure 9 presents scatter plots in the standard uniform space, for each category of piles, after applying the transformation given by the empirical distribution over the pile curve-fitting-parameters. There is an evident negative trend between the two curve fitting parameters for all categories of pile load tests. Thus, copulas that model negative dependences have to be employed.

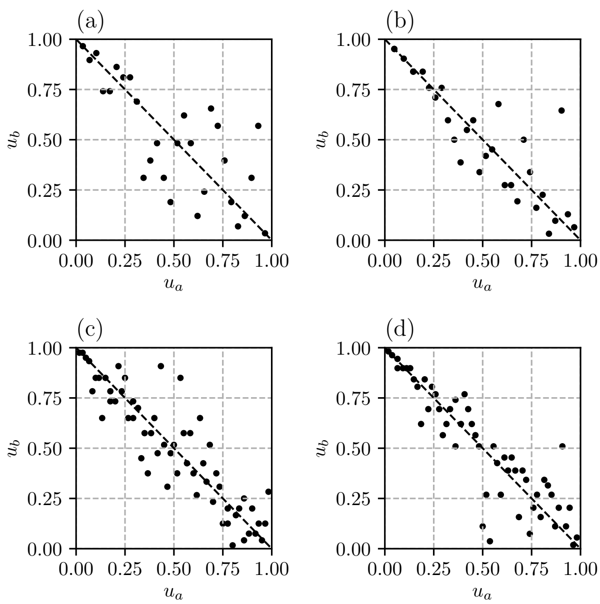


Figure 9 Scatter plots of pile curve fitting parameters from the dataset of Dithinde et al. [87] in the standard uniform space after applying the transformation given by the empirical distribution. (a) D-NC, (b) B-NC, (c) D-C, (d) B-C.

Dithinde et al. [87] demonstrated the inherent dependence between a and b parameters of equation (28), and employed a simple translation lognormal model to represent it, following the procedures described by Phoon et al. [94, 100] or Phoon [101, chapter 10]. On the other hand, Li et al. [95] modeled the dependence between a and b using a Gaussian copula. Later, Li et al. [37] conducted a comparison between not only two types of Gaussian copulas (the one adjusted to data through a direct computation of the Pearson's rho coefficient and the one fitted to data through Kendall's tau coefficient) but also with other three copulas capable of handling a wide range of negative dependence, namely Plackett, Frank, and Nelsen No. 16. They used a methodology to adjust copulas analogous to the one employed to fit copulas for soil shear strength parameters of section 5.1.

In this section, an analogous modeling to that of Li et al. [37] is performed in order to present some worth mentioning results and remarks, obtained from the dependence of the dataset of Dithinde et al. [87]. In addition to the original four families of copulas proposed by Li et al. [37], this work also includes the Student-t copula in the analysis, following the recommendations of Embrechts et al. [12]. Concerning the fitting parameters of the Gaussian copula, only the Gaussian copula based on Kendall's tau will be taken into account, and not the one based on a direct computation of Pearson's rho. As stated by Li et al. [95, 37], the Gaussian copula adjusted by Kendall's tau coefficient performs better than the Gaussian copula fitted by a directly computed from data Pearson's rho coefficient.

Table 11 presents the fitting parameters of the five different studied copulas. All the dependence parameters are defined through the method of moments by relating Kendall's tau and the copula parameter of dependence. When there is no analytical solution of the relation between Kendall's tau and the copula parameter of dependence, such as in the case of the Plackett copula, a numerical approximation is applied over equation (15). On the other hand, the second parameter of the Student-t copula corresponds to the number of degrees of freedom, which is defined through a maximum likelihood procedure using the available data.

Table 11 Fitting parameters of the studied copulas for the pile curve fitting parameters.

Pile Test	Gaussian θ	Student-t θ; ν	Plackett θ	Frank θ	Nelsen No 16 θ
D-NC	-0.806	-0.806; 1.97	0.048	-7.85	0.008
B-NC	-0.924	-0.924; 1.50	0.015	-14.139	0.002
D-C	-0.918	-0.918; 3.50	0.016	-13.512	0.003
B-C	-0.927	-0.927; 2.29	0.014	-14.471	0.002

A graphical contrast between the measured and the simulated data from all these copulas, for the driven piles in cohesive soils dataset, is shown in figure 5.2. Following Li et al. [95, 37], a lognormal marginal distributions are adopted for representing both a and b parameter. Similar scatter plots can be obtained for the other pile load test categories.

As stated in section 5.1, it is impossible to define the best-fit copula model by a simple visual inspection of the simulated data, such as in the scatter plots of figure 5.2. For instance, the best-fit copulas must be defined quantitatively based on some criteria. For example, Li et al. [95, 37] employed the widely applied AIC/BIC methodologies to determine the best-fit models. This methodology is also adopted in this document. Then, AIC and BIC coefficients are computed and presented in table 5.2 for all the studied copulas and all the pile tests categories.

From table 5.2, there can be extracted some important observations worth mentioning. Firstly, note that although the dataset is the same, there are some minor discrepancies between the AIC/BIC values reported in this document and the ones reported by Li et al. [37]. These minor discrepancies are due to the number of decimals employed for the computation of the AIC/BIC values. Nonetheless, these minor discrepancies do no have a relevant impact on the final results since the same best-fit copulas are obtained.

Second, observe in table 5.2 that AIC/BIC values in bold highlight the best-fit copulas. Although both selection criteria mostly

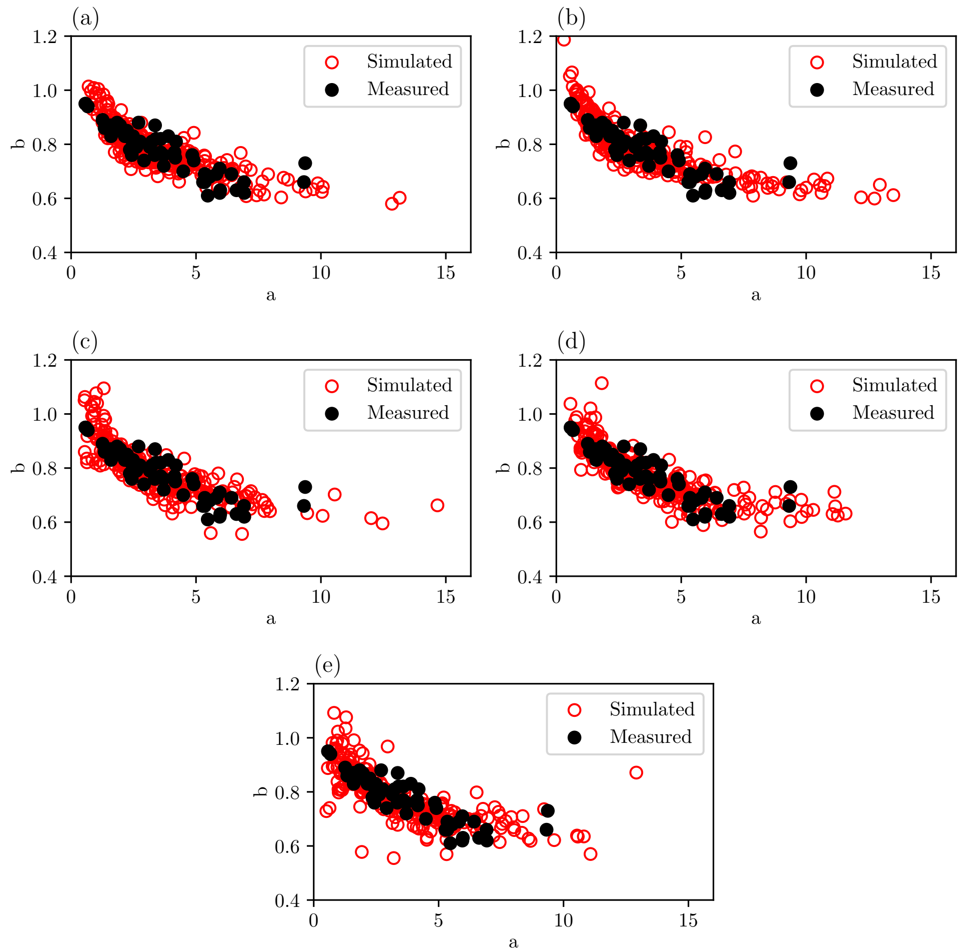


Figure 10 Simulation of the D-C dataset from different copulas using Lognormal marginal distributions. (a) Gaussian copula, (b) Student-t copula, (c) Plackett copula, (d) Frank copula, (e) Nelsen No 16 copula.

Table 12 AIC and BIC values for the different studied copulas.

Pile Test	Gaussian		Student-t		Plackett		Frank		Nelsen No 16	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
D-NC	-26.07	**-24.73**	**-26.33**	-23.67	-23.32	-21.99	-22.03	-20.70	-16.28	-14.95
B-NC	-42.00	-40.60	**-50.60**	-47.80	-49.98	**-48.58**	-42.99	-41.59	-44.23	-42.83
D-C	-81.64	-79.56	-86.93	-82.78	**-90.27**	**-88.20**	-90.04	-87.96	-83.62	-81.54
B-C	-73.94	-71.97	-87.83	-83.89	**-88.97**	**-87.00**	-81.79	-79.81	-84.83	-82.86

Values in bold highlight the best-fit copulas

agree in the determination of the best-fit copula, there are some cases where they disagree. Particularly, this situation is evident in the copula definition of the D-NC and B-NC datasets, and most specifically concerning the Student-t copula.

While the AIC defines the Student-t copula as the best-fit one for the D-NC and B-NC datasets, the BIC defines the Gaussian and the Plackett copulas, respectively. These discrepancies are due to the number of parameters of the Student-t copula, and to the penalty that the BIC assigns to it. BIC penalty for the number of model parameters is greater than the one of AIC. Then, taking into account that the Student-t copula is a two-parameter copula, its penalization in the BIC is higher. In this way, as the adjustment of the Student-t, Gaussian, and Plackett copulas are quite similar for these two datasets, a greater penalty may lead to changes in the best-fit copula selection, as it happens in this case.

Third, observe from table 5.2 that, for the same dataset, some AIC values from different copulas are quite close, likewise some BIC values. Specifically, this is evident for the Student-t copula and Gaussian copula from the D-NC dataset, Student-t copula and Plackett copula from the B-NC dataset, or the Plackett copula and Frank Copula from the D-C dataset. These close values indicate that both copulas may be suitable for modeling the dataset since their adjustment is quite similar. As stated by Li et al. [37], for the D-C dataset case where the Plackett copula is defined as the best-fit copula, Frank copula may be also a suitable copula for modeling the dependence between a and b parameters. In this

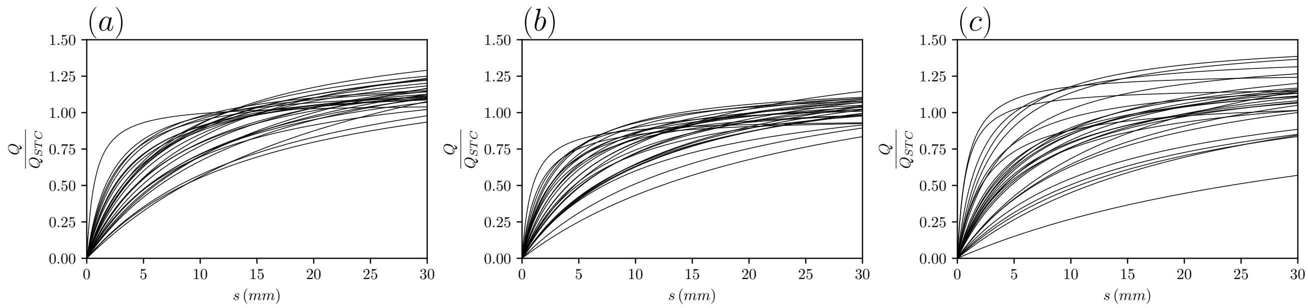


Figure 11 Hyperbolic curves of driven piles in noncohesive soils (D-NC). (a) measured, (b) simulated from a Gaussian copula, (c) simulated without dependence.

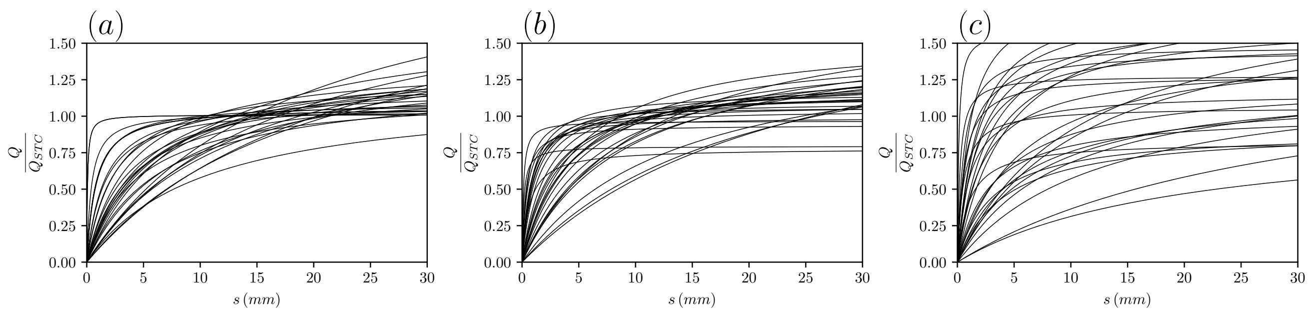


Figure 12 Hyperbolic curves of bored piles in noncohesive soils (B-NC). (a) measured, (b) simulated from a Plackett copula, (c) simulated without dependence

sense, when the discrepancies between AIC/BIC values of two copulas are quite small, no major differences are expected in their behavior when modeling dependence between random variables.

To be in concordance with Li et al. [37], the same best-fit copulas will be adopted in this work since this change does not have a major impact on the final curves.

Now, it is interesting to show a graphical comparison among the measured, simulated from the best-fit copula, and simulated without dependence, hyperbolic curves. Figure 11 shows the curves for the D-NC dataset, figure 12 for the B-NC dataset, figure 13 for the D-C dataset and figure 14 for the B-C dataset. Similar comparatives were performed by Dithinde et al. [87], Li et al. [95, 37].

Similarities between measured and simulated with dependence hyperbolic curves are evident. Unlike assuming independence between a and b parameters of equation (28), copula theory can adequately represent their dependence and capture the remaining scatter of the normalized hyperbolic curves. In this way, copula theory demonstrates to be a suitable tool when capturing model uncertainty of normalized pile load-displacement curves. This applicability could be evaluated and extended to other types of geotechnical models where there may be dependence between the fitted parameters.

Nonetheless, it is worth mentioning that since the work conducted by Dithinde et al. [87] the lognormal probability function has been adopted to represent both curve fitting parameters, a and b, without having evaluated other PDF. This assumption was replicated without validation by Li et al. [95], and later by Li et al. [37]. Therefore, other univariate probability distributions have not been evaluated till the moment to model these two parameters. In this regard, Li et al. [102] demonstrated that other univariate probability distributions, different from the lognormal, such as Weibull or Gumbel truncated below zero distributions, are more suitable to represent the two hyperbolic curve fitting parameters of equation (26) or (27) in most cases. In this sense, by evaluating other marginal distributions, simulated curves of the figures 11, 12, 13, and 14, can get a better approximation to the measured data.

5.3 Geotechnical earthquake parameters

Earthquakes are one of the most destructive forces in nature and have a direct impact on civil structures and human life, so it is important to adequately characterize their properties in order to consider them in designs. From the geotechnical perspective, several ground motion parameters (GMP) have been documented, studied, and measured. Examples are peak ground acceleration (PGA), peak ground velocity (PGV), pseudo-spectral acceleration (PSA), cumulative absolute velocity (CAV), among others [see 103, for more details].

Furthermore, it is common in practice to employ a ground motion prediction equation (GMPE) in order to predict peak ground motions parameters and response spectra of an earthquake, as functions of its magnitude, distance, and other properties. As noted by Cornell [104] and McGuire [105], GMPEs are critical elements when conducting probabilistic seismic hazard and risk analysis. Thus, quite a few GMPEs have been developed for

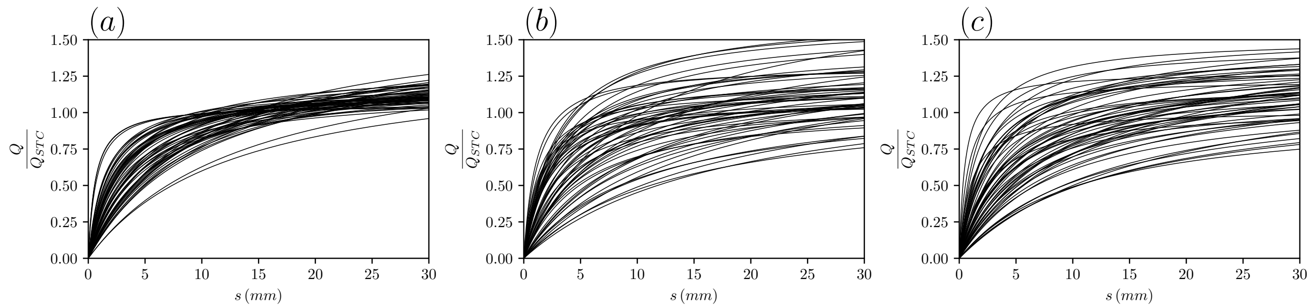


Figure 13 Hyperbolic curves of driven piles in cohesive soils (D-C). (a) measured, (b) simulated from a Plackett copula, (c) simulated without dependence

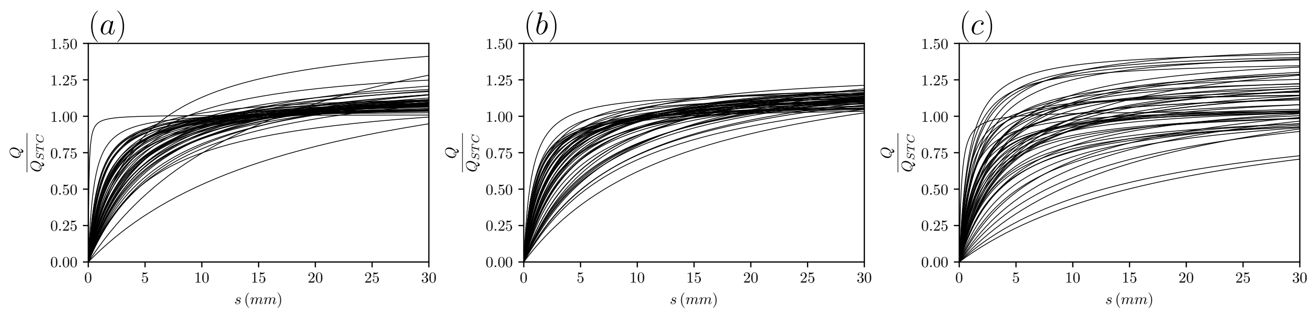


Figure 14 Hyperbolic curves of bored piles in cohesive soils (B-C). (a) measured, (b) simulated from a Plackett copula, (c) simulated without dependence

different seismic regions and different GMPs, such as the PGA, PSA, CAV, etc.

In practice, it is common to independently predict the GMPs using GMPEs, and to assume that the GMPs are lognormally distributed [106]. That is to say, from the same seismic event, each GMP has its respective equation and the resulting parameters are assumed to follow a lognormal probability distribution. Nonetheless, a single GMP is not sufficient to correctly represent the behavior of ground motions, so multiple GMPs are needed to capture different features of an earthquake [107]. Additionally, given that the GMPs have a common source, attenuation and site attributes, these GMPs are indeed interrelated, and therefore, cannot be modeled as independent entities [106].

In this sense, attempts have been made to characterize correlation among some GMPs. For this purpose, early works such as the one conducted by Baker and Cornell [108] have employed the popular linear correlation coefficient. The application of Pearson's rho correlation coefficient was to be expected, given its simplicity and popularity, so several works have also employed it. For instance, the linear correlation among the spectral acceleration (SA), Arias intensity (AI), and PGA, was studied by Baker [109], while the linear correlation between the SA values at different periods of vibration was studied by Baker and Jayaram [110], Goda and Hong [111] and Goda and Atkinson [112].

Disadvantages regarding linear correlation coefficients, and specifically Pearson's rho correlation coefficient, have already been exposed in this document. Furthermore, the multivariate normal distribution has been commonly employed for representing the dependence of multivariate ground-motion parameters [106]. From copula theory viewpoint, the multivariate normal model is not the only dependence structure available, and indeed it may not be the most adequate, so other dependence structures have to be assessed.

Although the importance of copula theory to model GMPs is evident, its use has been not so popular. One of the pioneering works was presented by Goda and Atkinson [106], where dependence between PGA and PSA parameters was studied for different vibration periods, employing copula theory. At the same time, Goda and Atkinson [106] validated the conventional approach, i.e., to fit lognormal distributions to the GMPs and to characterize their dependence through a linear correlation coefficient, by comparing it to the approach of copula theory. They employed a strong-ground-motion database, obtained from the Pacific Earthquake Engineering Research-Next Generation Attenuation of ground motions (PEER-NGA). Two elliptical copulas were initially considered, namely the Normal copula and the Student-*t* copula. [106] demonstrates through lognormal-probability-papers [20] and the AIC, that the lognormal probability distribution is suitable as the marginal distribution for both parameters, and also that the Normal copula is appropriate to represent their dependence. Furthermore, Goda and Atkinson [106] emphasizes in the benefits of using Kendall's tau coefficient to compute Pearson's rho through equation (14), instead of directly calculating Pearson's rho correlation coefficient. In this way, although the two-step approach was validated from the copula perspective, the benefits of using Kendall's tau coefficient are highlighted.

Another work in this framework was conducted by Xu et al. [107], in which the PGA and CAV ground motion parameters are studied and modeled as dependent parameters through copula theory. For this purpose, the Taiwan Strong Motion Instrumentation Program database was employed, and unlike Goda and Atkinson [106], several marginals distributions and copulas were evaluated. However, Xu et al. [107] came to a similar conclusion to that of Goda and Atkinson [106], since it was found that the lognormal distribution is suitable as the marginal distribution for the GMPs and their dependence can be represented through a Gaussian copula. Thus, both studies validate the common two-step approach for their respective GMPs and database.

Additionally, Xu et al. [107] also constructed a copula model subject to a given magnitude and distance from a seismic event, using GMPEs for both PGA and CAV parameters. Normalized residuals between the predicted PGA and CAV parameters subject to a given magnitude and distance were also modeled through copula theory. Again, Gaussian copula and the lognormal distribution were also found suitable to model residuals of the PGA and CAV parameters, which is in concordance with the commonly used two-step approach.

More recently, Cheng et al. [113] employed the vine copula theory to generate a multivariate joint probability function of ground pseudo-spectral accelerations at different vibrations periods. The advantages of using vine copulas for the model were evident. Vine copulas not only better capture the multivariate dependence, but also manages to capture tail dependence between pairs of variables. The methodology proposed by Cheng et al. [113] performs much better than the commonly employed multivariate normal distribution.

Finally, it is worth mentioning that the studies of ground-motion-parameters are characterized by having a substantial amount of data. It has already shown that the amount of data when studying shear strength parameters or settlement measures is commonly scarce; indeed, just on unusual occasions this amount of data exceeds a hundred. However, this is not the case of the GMPs, since in recent years the interest in earthquakes has grown by engineers, researchers, and especially, governments. For this reason, several programs around the world have been implemented to measure data related to GMPs in real-time, which allows having a large amount of information. One good example is the labor conducted by the *Pacific Earthquake Engineering Research Center*, where a large amount of geotechnical information (and of different kinds) on earthquakes has been compiled, analyzed and shared.

5.4 Soil spatial modeling

Variability soils and rocks is evident and cannot be neglected. In fact, their properties may vary significantly, even for the same deposit. This variability is derived from different internal and external agents such as the deposition, post-deposition, weathering processes, anthropic actions, etc [114]. Thus, as it is impossible to exactly know the soil or rock properties in every single point of a deposit (or domain, mathematically speaking), and there will be intrinsic uncertainties associated with their characterization. So, it becomes mandatory to model soil or rock spatial variability and predict their properties to obtain a better description of their behavior under different circumstances.

Several methodologies are found in the literature for characterizing the spatial variability of materials, in this case soils or rocks. Nonetheless, not all these methodologies include the dependence among soil-rock parameters in their predictions, so they assume independence or in the best case a Gaussian dependence structure. Most recently, the applicability of copula theory for modeling spatial variability with dependent parameters has been evaluated by some pioneering works. The present section aims to review these particular methodologies and the role that copulas have performed.

Random field theory [see, e.g., 115] constitutes one of the most comprehensive and popular methodologies applied for studying spatial fluctuation of material properties, such as the ones of soils and rocks. Succinctly, a *random field* is a function that defines random quantities of some variables over an arbitrary domain, which can be a multi-dimensional space. Thus, a random field is a function of the form $H(x,\omega)$, which associates a random quantity with a continuous index $x \in X$, and a coordinate ω in the sample space Ω. Note that $X \subset \mathbb{R}^d$, for $d \in \{1,2,3\}$, is a bounded domain that represents the physical space.

Regarding dependence in the theory of random fields, two main types of correlations among variables are defined, namely the *cross-correlation* and the *auto-correlation*. The former describes the relationship between values of two (or more) random variables at the same point of the physical domain X. The latter describes the relationship among values of one random variable at two different locations in the domain X. Accordingly, observe that based on these definitions the dependence between shear strength parameters modeled in section 5.1 corresponds to a cross-correlation dependence.

Commonly, cross-correlation in geotechnical random fields is either obviated or taken into account through a cross-correlation matrix M_c, where each term represent the correlation coefficient between the random variables. However, the importance of dependence has been elucidated in previous sections, so assuming independence may be misleading and lead to mistakes. Furthermore, cross-correlation matrices M_c have regularly implied the use of the Pearson's rho correlation coefficient and so, the multivariate Gaussian distribution. As noted by Zhu et al. [114] and Wang et al. [116], the independence assumption or a cross-correlation given by a Gaussian distribution leads to mistakes. So, it is necessary to evaluate different structures of dependence for modeling cross-correlated random fields, and copula theory shows to be suitable for this purpose.

After identifying these gaps, Zhu et al. [114] proposed a new methodology based on copula theory to deal with the cross-correlation dependence among variables in geotechnical random fields. In the first instance, this early work generates independent multivariate anisotropic random fields through auto-correlation coefficients ρ_s obtained by a formulation of an elliptical scale of fluctuation. These auto-correlation coefficients serve to create an auto-correlation matrix M_a, whereby some transformations, like the Cholesky decomposition, generate an independent random field.

Consequently, Zhu et al. [114] proposed some procedures to transform the originally independent random field into a cross-correlated random field that follows the dependence structure of a given copula. Specifically, they incorporated the dependence structure of the Gaussian, Student-t, Clayton, and Frank copulas into the cross-correlated random fields. The study performed a comparison between a copula-based cross-correlated random field of liquid limit, plasticity index and natural moisture content, and the original dataset of the same parameters. As a result of this comparison, Zhu et al. [114] validated the proposed methodology since the statistics of both agree and, after the calibration, the copula-based cross-correlated random field generates a good representation of reality.

Finally, Zhu et al. [114] also studied some incidences of the dependence parameter and type of selected copula over the final random fields. However, the scope this work is rather limited, so other studies complement it. Wang et al. [116] further studied the incidences of copulas over the probabilities of failure of a homogeneous slope, under different conditions, modeled through the copula-based cross-correlated random field proposed by Zhu et al. [114]. In this sense, Wang et al. [116] employs the subset simulation algorithm [117] to estimate the slope probabilities of failure by varying the copula selection, covariance of marginal distributions, the scale of fluctuation of the random field, and the cross-correlation coefficient.

Further details about the incidence of copulas in the final probability of failure of geotechnical structures are exposed in section 7. For the moment, what is truly important in this section is to highlight that cross-correlation plays a crucial role when modeling soil spatial properties.

Now, regarding the auto-correlation in the theory of geotechnical random fields, a new approach for its modeling based on copula theory was recently proposed by Wang and Li [118]. Note that auto-correlation is a different dependence concept from those studied so far in this document. Auto-correlation dependence is expected to be almost perfect for quite close points in the random field domain but simultaneously approaches zero for distant locations. In this sense, a copula parameter of dependence will be variable on the domain, and the selected copulas must be able to model from almost perfect dependences to almost null dependences.

The auto-correlation dependence structure of random fields has commonly been based on the assumption of gaussianity. However, this assumption is not always verified, and in the light of copula theory, Gaussian dependence is not the only option. Thus, Wang and Li [118] evaluated the use of different dependence structures for a one dimensional random field, supported on copula theory. To this end they employed several copulas, including vine copulas.

Wang and Li [118] emphasizes that the use of different dependence structures has a strong impact on the random field, except when this last one is highly ragged or highly smooth. Thus, a correct characterization and selection of the dependence structure (copula) is highly important when representing auto-correlation of a random field. Additionally, as the copula dependence parameters vary on the domain, candidate copulas to represent the auto-correlation dependence must be able to handle a wide range of dependence values, such as the comprehensive copulas [118].

On the other hand, there are some methodologies worth mentioning related to the approaches of geospatial interpolation, such as the kriging prediction method. The majority of these approaches are based on some assumptions, such as a Gaussian dependence structure given by a variogram. However, this particular assumption has to be validated since it may not be an accurate representation of reality and leads to mistakes, as stated before.

Some pioneering works have explored the use of copulas in this regard. For example, Marchant et al. [119] employed a Gaussian copula with generalized extreme value marginal distributions to spatially model observations of soil cadmium concentrations, whose dataset exhibits outliers and tails. On the other hand, Kazianka and Pilz [120] proposed three methodologies based on the kriging approach and copula theory to model a strongly skewed soil radioactivity measurements. They demonstrated that the proposed methodologies generalize some of those already existing in the state-of-the-art. Later, Kazianka and Pilz [121] employed a Bayesian spatial interpolation method, in conjunction with copula theory, to model the same dataset employed in Kazianka and Pilz [120], also exploring incidences of the uncertainties in fitting parameters.

There are other studies in the literature regarding the use of copulas in geospatial interpolation. An exposition of these works is out of the scope of this document, taking into account that their framework moves away from the geotechnical context. However, it is worth mentioning that a general observation of these studies is the transcendence of the dependence structure when interpolating geospatially. Outliers and skewness are common in geostatistics datasets, so their dependence structures are not correctly captured by the assumption of Gaussianity. Thus, it is necessary to evaluate other dependence structures, and copula theory is especially suitable for this end.

Finally, it is worth highlighting that the use of copulas in the modeling of soil spatial properties has not widely been studied, and further investigations must be carried out in this regard. Particularly, in the case of random fields, there are gaps in the state-of-the-art that can be explored through copula theory. For example, the auto-correlation in two or more dimensions considering copulas can be studied, or the implementation of the auto-correlation and cross-correlation in random fields, each one represented by copulas.

5.5 Other uses of copulas in geotechnics

Copulas have been mostly employed in geotechnical engineering to model some types of dependence such as the one of shear strength parameters (section 5.1), settlement parameters (section 5.2), GMP (section 5.3) or spatial interpolation (section 5.4). However, there have been other applications of copulas in geotechnics that have not been as widely explored, and that will be mentioned in the following.

5.5.1 Landslide hazard assessments

Motamedi and Liang [122] evaluated the applicability of copulas for performing landslides hazard assessments in a regional scale. These assessments are characterized by three elements, namely magnitude M, frequency T and location (also called susceptibility) S. Assuming that these three elements are independent variables, which is the common practice, the following equation for hazard assessments is obtained [122]:

$$H = P(M)P(T)P(S) \tag{29}$$

where H is the landslide hazard, $P(M)$ is the probability of a landslide magnitude, $P(T)$ is the probability of a landslide recurrence, and $P(S)$ is the susceptibility or spatial probability.

Using a recorded landslide dataset at the Western Seattle area, Motamedi and Liang [122] validated instead the following equation:

$$H = P(T)P(M \cap S)$$

in which T is modeled as an independent variable, but in turn M and S are modeled as dependent variables through the theory of copulas.

Motamedi and Liang [122] evaluated a total of 14 families of copulas for the modeling of the dependence between M as S. They found that for this dataset the Gumbel copula bests represents the structure of dependence between M and S. Furthermore, they performed a contrast between the results of the proposed methodology and the ones from equation (29), thus finding that the proposed approach makes a better representation of reality. Dependence, and hence copula theory, is suitable and even necessary in landslide hazard assessment on a regional scale [122].

5.5.2 Earthquake early warning

Earthquake early warning (EEW) methodologies aim to detect earthquakes as soon as possible in order to disseminate alerts and prevent people before the peak ground motion arrival. For this purpose, the relationship between soil early motions (SEM) and peak ground motions (PGM) must be defined.

Nonetheless, there are uncertainties in this relationship, which lead to missed or false alarms. Missed alarms are defined as the event of SEM less than a prescribed threshold, but PGM greater than a triggering threshold. On the other hand, false alarms are defined as the event of SEM greater than a prescribed threshold, while the PGM is smaller than a triggering threshold.

Looking to model these uncertainties, Wang et al. [123] employed a dataset of peak ground displacements within the first three seconds after P-waves arrives (as SEM parameter) and peak ground velocities (as PGM parameter), which are hereinafter referred to as PD_3 and PGV, respectively. They evaluated six copula function to model the dependence between these two parameters, and found in the Frank copula its best representation. Thus, this copula was calibrated from the dataset, and it was used to quantify the probability of missed and false alarms.

This copula model was compared with the conventional approach, which assumes independence between the parameters, and a better fit to the dataset and better results were found. In this way, taking into account the dependence between PD_3 and PGV, Wang et al. [123] showed that the probabilistic model of these two variables can be better characterized.

Additionally, Wang et al. [123] proposed a copula-based methodology to optimize the selection of the triggering PD_3, thus reducing the probability of missed alarms (PGV greater than the threshold) and false alarms (PGV less than the threshold). It is worth noting that, although it has not been yet done, this methodology could be extended to other types of geotechnical early warnings, such as the ones of debris flows or landslides.

5.5.3 Studies on sample size

Tang et al. [75] studied the impact of the sample size on the uncertainties in the definition of probabilistic geotechnical models. They assumed a true probabilistic model, which consists of a bivariate copula with two marginal distributions. From this model several samples of different size were drawn. For each sample size, a set of marginal distributions and copulas are fitted to data and the best-fit marginal/copula functions were defined through the AIC methodology. Variation of the fitting parameters of each model are studied, and since the true model is known, it is possible to compare it with those models defined on each sample size.

As expected, the larger sample size the greater the probability of selecting the true copulas, marginal distributions, and their parameters of adjustment. Furthermore, Tang et al. [75] also found that the correlation coefficient has a great impact on the selection of the correct copula model. The stronger the dependence, the smaller the amount of data needed to correctly select true copula model, and vice versa, i.e. the weaker the dependence, the larger amount of data will be needed.

Furthermore, Tang et al. [75] pointed out that another useful way to select the best copula and marginal distributions, for a dataset, is by comparing the variation of their AIC scores. As they pointed out, the best copula and best marginal distributions have less variation in their AIC scores. In order to obtain the AIC scores of the evaluated probabilistic models, and hence their variation, Tang et al. [75] recommended the use of the Bootstrap method [see, e.g., 124].

6 Uncertainties in the construction and definition of copula for geotechnical variables

It is quite unusual in geotechnical engineering to have a large amount of data to characterize soil and rock materials; thus, it is theoretically impossible to determine true probabilistic models. For this reason, in many cases, the statistics and models obtained from a sample, usually small, are assumed as obtained from populations. This assumption leads to uncertainties in modeling, which will be a function of the quantity and quality of information. These uncertainties have a direct impact on the veracity of the model and its results, such as the probability of failure.

Dealing with these uncertainties is not a trivial task, specially when the information is scarce. In any case, when it is not possible to reduce uncertainty, by collecting more and better information, a good practice is to characterize this uncertainty. Several investigations have been carried out with the aim of characterizing uncertainty when constructing and selecting copula models that represent dependence among geotechnical random variables. An example of these methodologies is the work of

Tang et al. [75], who studied of the impact of sample size in the copula model definition.

In this section, we will study the work of Li et al. [102], who proposes an interesting methodology for characterizing model uncertainty in the construction and definition of multivariate probability models when the information is scarce. This methodology focuses on the characterization of uncertainty for the different probabilistic models employed by Li et al. [37] for representing the pile load-displacement datasets of Dithinde et al. [87]. Recall that the works of Dithinde et al. [87] and Li et al. [37] were already introduced in section 5.2. Furthermore, it is worth mentioning that the procedures and remarks of this section can be extended to other examples of dependence of geotechnical random variables when the information is scarse.

For this purpose, the bootstrap method [124] was employed by Li et al. [102] for studying uncertainties in both copulas and marginal distributions. Bootstrap method belongs to the so-called *resampling techniques*, which are useful for deriving sampling distributions of statistics from the measured data.

Briefly, the idea behind the bootstrap method consists of creating N_s sets of bootstrap samples with the same sample size N of the original dataset. Each bootstrap sample is constructed by random sampling with replacements from the original dataset; that is to say, every single observation can appear once, more than once, or not at all. For each bootstrap sample, estimates such as the mean, standard deviation, copula parameter of dependence, and even AIC/BIC values, are computed. After obtaining all values of the desired measure, every estimative can be associated with a sampling distribution, which represents its uncertainty. Readers are referred to Most and Knabe [125] and Luo et al. [126] for additional applications of the bootstrap method in geotechnical engineering.

According to Li et al. [102], there are two kinds of uncertainty when constructing copula models: one related to the adjustment of the parameters of the marginal distributions and copulas, and another related to the definition of the best-fit marginal distribution and copula function. Li et al. [102] employed the bootstrap method to characterize the uncertainty of the sample statistics of the marginal distributions for different pile load-displacement databases, i.e. mean and standard deviation. Furthermore, this work also characterizes, through the bootstrap method, the uncertainty in the AIC scores when defining the best-fit models for both marginal distributions and copulas.

Uncertainty in the copula dependence parameter was mentioned, but not explicitly studied, by Li et al. [102]. Therefore, this uncertainty will be explored in this document for illustration of the bootstrap procedure and in order to complement the original work of Li et al. [102]. In this way, the database of Dithinde et al. [87] is adopted for adjusting a set of copulas and to conduct a bootstrap procedure in order to characterize the uncertainty in the definition of the fitting parameters of these copulas. The copulas to be used in this procedure will be the Gaussian, Plackett, Frank and Nelsen No. 16.

Note that the Student-t copula is not included in the analysis since, by employing the method of moments for the adjustment, its parameter of dependence is the same as the Gaussian copula. Therefore, results obtained for the Student-t copula would be the same of those of the Gaussian copula.

As recommended by Most and Knabe [125] and Luo et al. [126], $N_s = 10000$ is adopted in the bootstrap procedure conducted for the evaluation of the uncertainties in the dependence parameter of each copula. For each bootstrap sample, copula dependence parameters are computed by the method of moments, i.e., employing respective Kendall's tau coefficient and equation (15) or the corresponding simplifications exposed in section 4.4. So, at the end of the procedure there will be 10000 estimations of the dependence parameters to construct every sampling distribution.

Resulting sampling distributions for the parameters of dependence of this example can be observed in figure 6, where each graph corresponds to a copula, and each curve to a pile test category. Statistics from the estimated distributions are computed and summarized in table 6.

From figure 6 and table 6 some important remarks can be extracted. Firstly, by comparing tables 11 and 6, it is concluded that the bootstrap mean estimates of copula parameters of dependence are quite similar to those obtained from the original dataset. This fact indicates that distributions obtained through the bootstrap method can appropriately capture the characteristics of the data set.

Additionally, it is evident that the larger the dataset, the lower the dispersion of the models. This fact is noticeable by comparing, in figure 6, the curves of the D-NC data set (28 samples), with the ones of the D-C dataset (59 samples). Furthermore, these comparisons are quantified in table 6 where the coefficient of variation C.O.V. of the sampling distributions is higher when the dataset is smaller. Thus, the highest C.O.V. corresponds to the D-NC dataset, and the B-NC, B-C, and D-C data sets follow in order, which coincides with the amount of information of each pile load-displacement test. Therefore, this fact constitutes a motivation to collect more information, in order to obtain more accurate results.

On the other hand, uncertainty when selecting the best-fit marginal/copula models was also studied by Li et al. [102] through the computation of the AIC values for each copula in each pile dataset. It is interesting summarize some of these results here. Again, using $N_s = 10000$, a bootstrap procedure over the dataset of Dithinde et al. [87] is conducted. For each bootstrap sample, AIC scores for each copula are computed and the corresponding best-fit copula is defined. So, there will be 10000 AIC scores for each copula and 10000 best-fit copulas for each load-displacement test at the end of the procedure.

Figure 6 presents the sampling distribution of the AIC values for different copulas and pile tests. Additionally, table 14 contains the number of times that each copula is identified as the best-fit copula for each pile test.

Figure 6 demonstrates that the bootstrap mean estimates of the AIC scores match well with those values obtained from the original dataset. However, AIC bootstrap sampling distributions for each copula are broad, which means that there may be large uncertainties when identifying the best-fit copula to represent the dependence between a and b parameters. Again, uncertainties (which are related to coefficients of variation C.O.V. values) are a function of the amount of data. Thus, the more available

Figure 15 Sampling distributions for the parameters of dependence of different copulas and for the different pile tests of Dithinde et al. [87]. (a) Gaussian copula, (b) Plackett copula, (c) Frank copula, and (d) Nelsen No 16 Copula

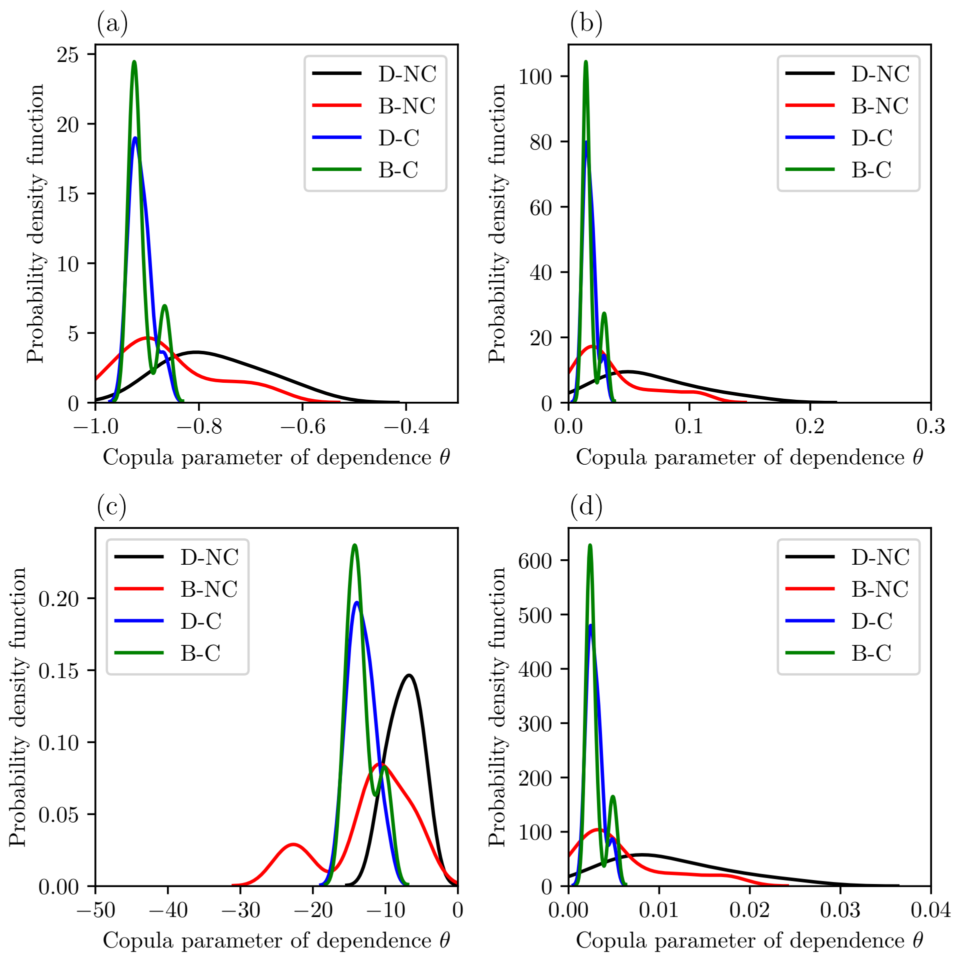


Table 13 Mean and coefficient of variation C.O.V. from the obtained sampling distributions for the copula dependence parameters in the different pile tests of Dithinde et al. [87].

Pile Test	N	Gaussian		Plackett		Frank		No 16	
		Mean	C.O.V.	Mean	C.O.V.	Mean	C.O.V.	Mean	C.O.V.
D-NC	28	-0.795	0.135	0.068	0.797	-8.643	0.395	0.010	0.842
B-NC	30	-0.914	0.068	0.029	0.961	-16.720	0.519	0.003	0.992
D-C	59	-0.916	0.030	0.017	0.418	-13.964	0.200	0.003	0.402
B-C	53	-0.923	0.040	0.020	0.583	-15.408	0.291	0.003	0.600

Table 14 Number of times that each copula is defined as the best fit copula for each pile test category

Pile test	Gaussian	Student-*t*	Plackett	Frank	No 16
D-NC	**5511**	3610	750	48	81
B-NC	1600	3216	**4204**	364	616
D-C	861	1470	3327	**4207**	135
B-C	390	2188	**5589**	222	1611

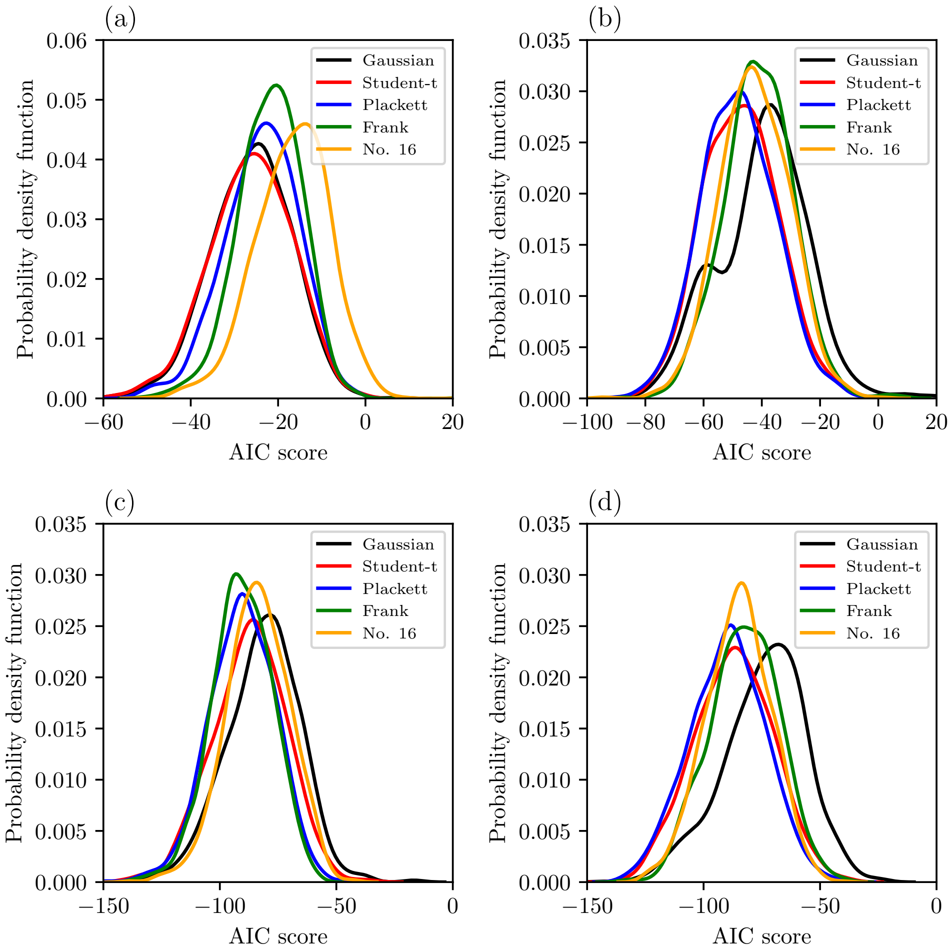
Values in bold highlight the copulas selected the most times as best-fit copulas

information, the smaller coefficient of variation and the lower uncertainty.

Table 14 shows that no copula can be identified as the best-fit copula with total certainty, that is to say, with a probability of one or almost one. There are copulas with a higher probability of being the best-fit copula, such as the Gaussian in the D-NC pile test, the Plackett in the B-NC and B-C pile tests, or the Frank in the D-C pile test. However, there is no total certainty in the copula selection, and therefore this uncertainty must be considered in the models.

Additionally, note from table 14 that copulas with a major probability of being the best-fit copula mostly agree with those defined in table 5.2. However, interestingly this is not the case of the D-C pile dataset, where Frank copula has the major probability of being the best-fit copula. Li et al. [37] pointed out that the Frank copula, instead of the Plackett copula, may also be suitable to

Figure 16 AIC probability density functions of different copulas for different piles. (a) D-NC dataset, (b) B-NC dataset, (c) D-C dataset, and (d) B-C dataset



model the D-C dataset, and this fact is verified by this bootstrap procedure of the present section. Thus, the bootstrap method also demonstrates to be quite useful to define best-fit copula by taking into account model selection uncertainties.

Li et al. [102] emphasizes the importance of the bootstrap method to perform reliable judgments when selecting best-fit copulas, especially when the sample size is small, and therefore the uncertainties large. Uncertainties must be incorporated on the selection of the best-fit copulas and the estimation of model parameters. Hence, the bootstrap method proves to be an adequate tool.

Finally, it is worth mentioning again that although the work conducted by Li et al. [102] is applied to the database collected by Dithinde et al. [87], it could also be extended to other geotechnical parameters. Concepts of model uncertainties are quite general, and so also the results presented in this section.

7 Reliability analysis of geotechnical models considering dependence through copula theory

Copulas have been employed in the reliability analysis of different geotechnical structures and models. Some examples are infinite slopes [82, 83], slopes with a circular failure mechanism [74, 76], foundations [82, 77], rock wedge slopes [80], retaining walls [85, 35, 80], among others. However, even though all these studies employ copulas, and indeed have several similarities, they have different methodologies and scopes that nurture the state-of-the-art.

This section presents the most crucial remarks and findings in this framework so far. Some examples and graphs will be replicated to highlight some cornerstones in the reliability analysis using copulas in geotechnical engineering.

Let us recall that a probability of failure P_f, can be understood as the probability of exceeding a given threshold or limit state, which leads to unwanted behaviors or instability. Mathematically speaking, the probability of failure is defined as the probability measure of a failure domain $\{x \in \mathbb{R}^d : g(x) \leq 0\}$, which is delimited by a limit state function $g : \mathbb{R}^d \to \mathbb{R}$. Some examples of limit state functions are: $g(x) = FS(x) - 1.0$, where FS is the factor of safety, or $g(x) = R(x) - S(x)$, where R are the resistance forces and S the applied forces.

In a general case, the probability of failure can expressed through the Stieltjes integral:

$$P_f = \int_{\mathbb{R}^d} \mathbb{I}[g(x) \le 0]\, \mathrm{d}F_X(x);$$

where F_X is the joint CDF of the vector of random variables $x \in \mathbb{R}^d$; when the associated joint PDF f_X exists, this last expression reduces to

$$P_f = \int_{\mathbb{R}^d} \mathbb{I}[g(x) \le 0] f_X(x)\, \mathrm{d}x; \tag{30}$$

here $\mathbb{I}$ represents the indicator function.

In a bivariate case, and considering that the random variables of the model are the shear strength parameters of the Mohr-Coulomb failure criterion, i.e., cohesion and friction angle, the probability of failure of the geotechnical model can be expressed as:

$$P_f = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{I}[g(c,\phi) \le 0] f(c,\phi)\, \mathrm{d}c\, \mathrm{d}\phi$$

where $f(c,\phi)$ is the bivariate probability density function of cohesion and friction angle.

7.1 The copula based sampling and copula direct integration methods.

In order to conduct reliability analysis, considering dependence through copula theory, two major procedures have to be conducted. First, it is necessary to define the random variables involved in the geotechnical system, evaluate their dependence, and construct the multivariate probability distribution using copulas. Details of these procedures are shown in section 4, and examples are presented in section 5. Second, a methodology must be employed to compute the probability of failure of the geotechnical model in order to assess its reliability.

There exist several methodologies in the literature to address this second step. Some of them are direct integration, transformations, or simulations. Each one of these alternatives has its particularities, advantages, and disadvantages. For instance, direct integration is accurate but may be unfeasible in very high dimensions or in very complex models, meanwhile simulation can be employed for complex and high dimensional models, but may be too computationally expensive when the probability of failure is low, and inaccurate when the number of simulations is not enough.

In the state of the art, two main techniques are used to estimate the probability of failure of geotechnical models when dependence through copula theory is considered: the direct integration method and the Monte-Carlo simulation. It is worth mentioning that these techniques have been renamed by some works as the *copula direct integration method (CDIM)* [see, e.g., 85, 73, 74] and the *copula based sampling method (CBSM)* [see, e.g., 76, 77, 127], respectively.

7.1.1 Direct integration method in copulas

When the random variables of the model are just soil cohesion c and friction angle ϕ, and its structure of dependence is given by a copula $\mathtt{C}$, equation (30) can be expressed through the following Riemann–Stieltjes integral, as:

$$P_f = \iint\limits_{[0,1]^2} \mathbb{I}\left[g\left(F_c^{(-1)}(u), F_\phi^{(-1)}(v)\right) \le 0\right] \mathrm{d}\mathtt{C}(u,v), \tag{31}$$

where $F_c^{(-1)}$ and $F_\phi^{(-1)}$ denote the pseudo inverses of the CDFs of c and ϕ, respectively.

If $\mathtt{C}$ is sufficiently differentiable and the copula density function $\mathtt{c}$ exists, then equation (3) can be replaced in equation (31), and hence, the latter can be rewritten as:

$$P_f = \iint\limits_{\mathbb{R}^2} \mathbb{I}[g(c,\phi) \le 0] f_c(c) f_\phi(\phi)\, \mathtt{c}\big(F_c(c), F_\phi(\phi)\big)\, \mathrm{d}c\, \mathrm{d}\phi. \tag{32}$$

This methodology was initially applied in geotechnics by Tang et al. [85] for the reliability analysis of an infinite slope and a retaining wall. Subsequently, other works such as the ones conducted by Phoon and Ching [35], Tang et al. [73] or Xu et al. [74] have employed it, also in a geotechnical context. Its concept is quite straightforward and can be applied in many geotechnical models. However, the direct integration method has some limitations related to the complexity of the model or the number of random variables.

Although equation (32) expresses exactly the probability of failure, its evaluation might be complicated, specially in high dimensions ($d \ge 5$). In that case, it is better to resort to simulation methods.

7.1.2 Monte Carlo Simulation in copulas

Monte Carlo Simulation (MCS) aims to approximate equation (30) through a repeated random sampling. Specifically, MCS draws samples from the copula that represents the dependence structure of the model, after that, these samples are transformed to the space of the basic variables through some method for estimating the inverse of the CDF of the marginal distributions. Then, the performance function g is evaluated for each one of these samples, and finally the probability of failure is computed as the ratio between the number of samples that leads to failure over the total number of samples.

Some researches, have highlighted the statistical error due to the number of samples underlying MCS in geotechnical reliability analyses [e.g., 85, 35]. This statistical error is aggravated in too small probabilities of failure since more samples are needed for accurate calculations. In this regard, it is true that the lower the probability of failure, the larger the sample size must be to avoid inaccuracies, and this may be time consuming.

This issue has been subject to an extensive discussion by different researchers since the number of samples of a MCS considerably affects the accuracy of the reliability analyses. In this regard, there exists several methodologies and equations to estimate the minimum sample size in a MCS. For example, equation

$$N_{min} \approx \frac{100}{P_f} \tag{33}$$

where N_{min} is the minimum sample size and P_f the expected probability of failure, is one of these estimates, and it is particularly one of the most straightforward to use.

It is beyond the scope of this document to further discuss the minimum sample size of a MCS and the repercussions of insufficient samples. So, readers interested in this topic are referred to Ang and Tang [128], and the references therein cited, for a more comprehensive introduction and details. Additionally, it is worth mentioning that there are currently other methodologies of simulation especially useful for too small probabilities of failure, such as subset simulation [see 117] or line sampling [129]. Some pioneering works have already integrated some of these methods of simulation on the reliability analyses of geotechnical structures that employ copulas [see e.g. 116].

Despite this fact, the MCS constitutes a straightforward method with many advantages, even over the direct integration method exposed by Tang et al. [85]. The MCS is perhaps one of the best current alternatives to conduct reliability analysis in the framework of geotechnical engineering when considering dependence through copula theory. The MCS does not need a linearization at the design point such as the FORM and SORM, neither an approximation to the limit state function for complex geotechnical problems such as the direct integration. Additionally, it can handle systems with many random variables, i.e., geotechnical systems of high dimensions.

From our point of view, the MCS is currently a very good alternative to conduct reliability analyses of geotechnical models where the dependence is considered. However, this methodology may be computationally expensive, especially for high dimensions and fairly small probabilities of failure. More advanced methodologies, such as subset simulation [see 117] or line sampling [130], exists. Alvarez et al. [131] presents a framework that allows to use subset simulation with copulas and random set theory, in order to conduct efficient reliability analysis where dependence and both aleatory and epistemic uncertainties are considered. Particularly in geotechnical engineering, Wang et al. [116] employed copulas in conjunction with subset simulation, in order to conduct reliability analysis of geotechnical structures.

7.2 Results of reliability analysis

Copulas have been used to estimate the probability of failure of several geotechnical structures. So, it is important to summarize some results and remarks, which may help us to understand how copulas impact reliability analysis of geotechnical models.

For this purpose, the infinite slope example studied by Tang et al. [85] is replicated in this document. This model is selected given its simplicity and that the incidence of copulas in the probability of failure has already been studied by Tang et al. [85, 73] under different conditions.

Similar outcomes and conclusions, from this infinite slope example, are obtained from other geotechnical structures, such as retaining walls [35, 85], slopes with a circular failure mechanism [76, 74] and three dimensional slopes [127], when there are more than two dependent random variables [77], or when several failure modes are considered at the same time [80].

Figure 17 presents a scheme of the infinite slope model to be analyzed. Here, H stands for the thickness of the unstable layer, γ represents the soil unit weight, α is the slope angle relative to the horizontal plane, and c and ϕ are soil cohesion and friction angle, respectively. Note that the groundwater table is assumed to be under the failure surface, and is not considered in the present analysis. Additionally, neither overload nor seismic loads are considered in this slope model.

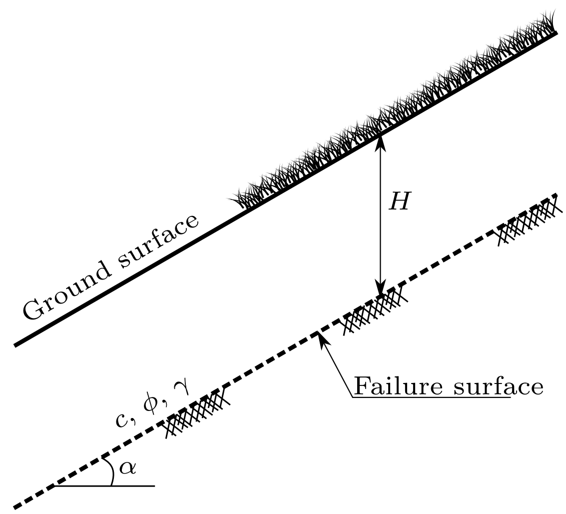


Figure 17 Infinite slope model

The factor of safety of the infinite homogeneous slope shown in figure 17 is given by

$$FS(c,\phi) = \frac{c + \gamma H \cos^2\alpha \tan\phi}{\gamma H \sin\alpha \cos\alpha} \tag{34}$$

The performance function for carrying out the reliability analysis will be of the form $g(c,\phi) = FS(c,\phi) - 1$, where the FS is the factor of safety obtained from equation (34).

In equation (34), both c and ϕ will be taken as random variables, and their dependence will be modeled through a copula. The other parameters of equation (34) are adopted as deterministic.

We will do three modifications to the original work of Tang et al. [85] related to the copula dependence parameter, the procedure to compute the probability of failure, and the evaluation of an additional copula. Pearson's rho correlation coefficient was employed by Tang et al. [85, 73] to estimate the dependence parameter of the different copulas used in the reliability analysis. However, in this document Kendall's tau coefficient is employed, given that it is a more suitable measure to represent the dependence in copula theory (see section 4.2). On the other hand, the direct integration was employed in the original example to compute the probability of failure. Nonetheless, the MCS will be employed for this purpose in this document, since it is a more general methodology, and it is also interesting to explore its performance and compare it with direct integration. Finally, in addition to the four copulas proposed by Tang et al. [85], the Student-t copula is evaluated, following the recommendations of Embrechts et al. [12]. In this way, the Gaussian, Student-t, Plackett, Frank, and Nelsen No. 16 copulas, will be employed in this example.

Regarding the other aspects from the original example of Tang et al. [85], these will be addressed without modifications. Thus, the lognormal probability distribution is adopted to model both c and ϕ, where the mean and coefficient of variation C.O.V. of c are taken as 12 kPa and 0.4 respectively, and the mean and coefficient of variation C.O.V. of ϕ are taken as 30° and 0.2 respectively. On the other hand, the deterministic parameters are taken as $H = 5$ m, $\alpha = 30°$ and $\gamma = 17$ kN/m^3. Finally, the Kendall's tau correlation coefficient τ between c and ϕ is -0.5.

Following the procedure of [85], the incidence of copulas in the probability of failure is studied under three conditions, namely: (1) geometry (H and α), (2) C.O.V. scaling factor of the marginal probability distributions ($C.O.V./\lambda_{c,\phi}$) and (3) correlation between c and ϕ ($\tau_{c,\phi}$).

Changes in each condition will be represented as changes in the so-called *nominal factor of safety* FS_n [85] (see e.g. Tang et al. [85] or Phoon and Ching [35]). The nominal factor of safety corresponds to the 5% fractile of all the obtained FS in a MCS. More details about FS_n, and the algorithm to compute it, can be found in the aforementioned two references.

The nominal factor of safety has the advantage of taking into account the coefficients of variation of the distributions of the shear strength parameters, and also the correlation between these two variables. This fact contrasts with a simple factor of safety, where only the mean values of the random variables are employed, so there is no room for uncertainty in the estimates of the factor of safety.

Unlike Tang et al. [85], in this document nominal factors of safety will vary between 1.0 and 1.16. That is to say, when plotting FS_n against the probability of failure of each copula, the horizontal axes (FS_n) for all the conditions are the same (i.e., $FS_n \in [1.0, 1.16]$), and their values are due variations of each particular condition. Specifically, $H \in [3.84\text{m}, 6.27\text{m}]$, $\alpha \in [27.7°, 32°]$, $\lambda_{c,\phi} \in [0.67, 1.64]$ and $\tau_{c,\phi} \in [-0.78, -0.26]$, and all these variations will lead to the same horizontal axis of FS_n.

Additional to the four copulas initially evaluated by Tang et al. [85], namely Gaussian, Plackett, Frank and Nelsen No. 16 copulas, Student-t copula is also studied in this document. Note that the aforementioned copulas are able to be adjusted to the Kendall's tau coefficient assumed in this example.

Then, MCS is performed for each copula, thus defining the probabilities of failure for aforementioned conditions. A total number of 10^7 samples will be simulated and employed by each MCS. The above number is not arbitrary since, from Tang et al. [85], the orders of magnitude of the probabilities of failure of the infinite slope example are known. So, by applying equation (33), a minimum sample size of 10^7 is obtained for the smallest probabilities of failure.

Note that the same sample size, equal to 10^7, is employed in this document for computing all the probabilities of failure, regardless of their magnitude. However, it would be possible, and indeed necessary, to adopt different sample sizes for the different orders of magnitude of the probabilities of failure. This practice would avoid excessive calculation times, in the cases that accurate estimates can be obtained with fewer samples. Nonetheless, the same sample size is employed in this document since, for this simple infinite slope model, it does not represent a problem for the current computer capacities.

Figure 7.2 condenses all the computed probabilities of failure, after applying the above procedures for all the studied copulas and conditions. In figure 7.2, each plot corresponds to the variation of a particular parameter, and each curve corresponds to a copula.

Additionally, table 7.2 presents a comparison of the probabilities of failure obtained through all the copulas, for all the studied failure mechanisms. For constructing table 7.2, probabilities of failure from the Gaussian copula are taken as reference, and the ratio $P_{f_i}/P_{f_{Gaussian}}$ is computed, in which P_{f_i} is the probability of failure obtained from each respective copula. Thus, all computed ratios are summarized in table 7.2.

In the first instance, it is worth mentioning that the probabilities of failure found in this document are in the same order of magnitude to those obtained by Tang et al. [85], but they are not equal. There are minor discrepancies due to the modifications applied; specifically, the use of Kendall's tau coefficient instead of Pearson's rho coefficient to compute the copulas dependence parameter contributes significantly to these differences. It is emphasized again that the use of Pearson's rho coefficient must be avoided as a measure to characterize the dependence in copula theory. Rank measures of dependence are more suitable for this purpose. Furthermore, when the dataset is available, the methods exposed in section 4.4 are also appropriated to fit copulas to data.

Nonetheless, it is worth noting that for the same dataset, and for the same copula, methods exposed in section 4.4 may also lead to a different copula dependence parameters, although quite similar. These minor discrepancies in the dependence parameter of a given copula will result in different probabilities of failure. Further studies are necessary to explore the repercussions of the copula fitting methods on the probability of failure of geotechnical structures.

Regarding the results presented in figure 7.2 and table 7.2, it is important to present some conclusions that help to understand the implications of using copulas when performing reliability analysis in geotechnical engineering. More details and remarks in this regard are presented by Tang et al. [85] and Tang et al. [73] for the infinite slope example, or by Phoon and Ching [35], Xu et al. [74], Li et al. [80] and Xu and Zhou [127] over other structures and conditions. Thus:

- As it was expected, as H or α decreases, the probability of failure becomes smaller, and consequently, the factor of safety increases. This conclusion is easily extendable to other geotechnical models by saying that as the geometric parameters become more adverse for the stability of the geotechnical system, the probability of failure will be greater, and vice versa.

Figure 18 Probabilities of failure associated with different copulas for an infinite slope analysis. (a) H varying from 6.27 to 3.84, (b) α varying from 32° to 27.7°, (c) $\lambda_{c,\phi}$ varying from 0.67 to 1.64, and (d) $\tau_{c,\phi}$ varying from −0.26 to −0.78

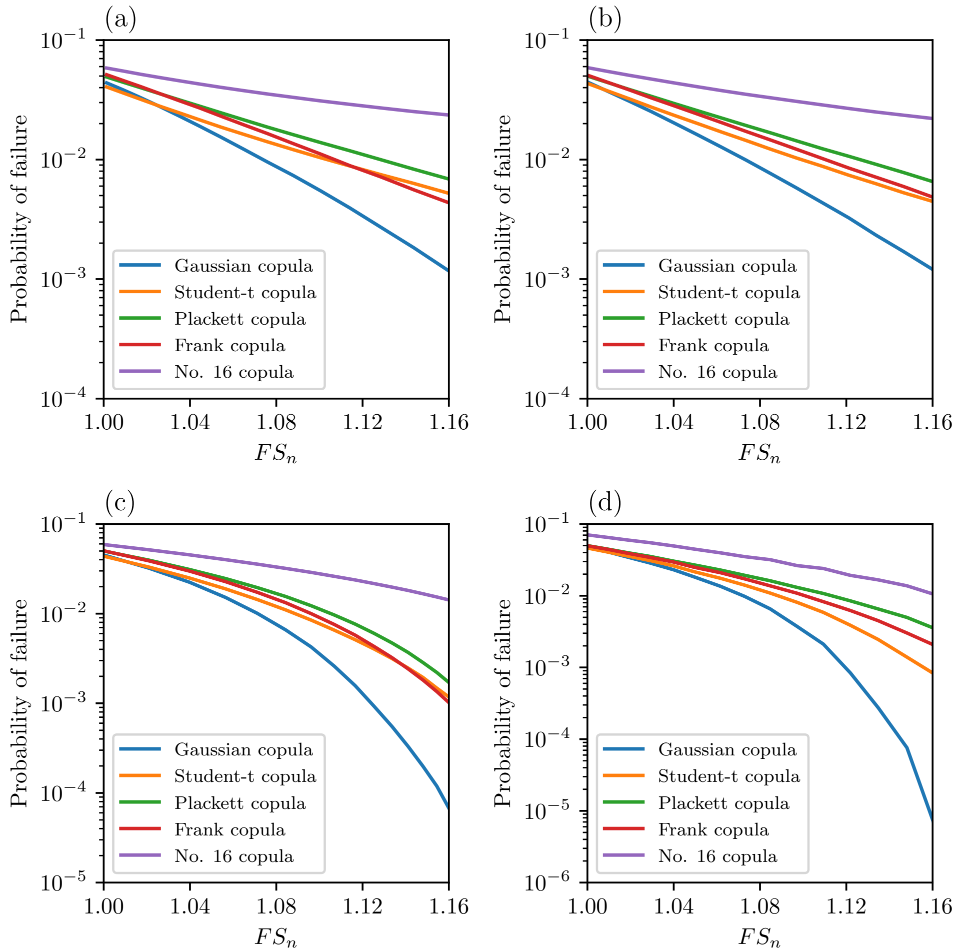


Table 15 Comparison of the P_f obtained from different copulas for the infinite slope example, and taking the P_f of the Gaussian copula as reference for computing the ratio $P_{f_i}/P_{f_{Gaussian}}$.

Copula	$H=[6.27,3.84]$			$\alpha=[32,27.7]$			$\lambda=[0.67,1.64]$			$\tau=[-0.26,-0.78]$		
	1.00	1.08	1.16	1.00	1.08	1.16	1.00	1.08	1.16	1.00	1.08	1.16
Gaussian	1	1	1	1	1	1	1	1	1	1	1	1
Student-t	0.94	1.69	4.35	0.98	1.57	3.78	0.98	1.69	21.37	0.98	1.68	90.00
Plackett	1.13	2.26	5.81	1.12	2.17	5.52	1.11	2.40	29.02	1.06	2.56	395.89
Frank	1.18	1.89	3.73	1.13	1.88	4.05	1.11	2.00	17.20	1.06	2.18	215.11
No 16	1.33	4.51	19.15	1.32	4.27	18.48	1.29	4.79	236.85	1.49	4.97	1144.33

- As the coefficient of variation C.O.V. decreases (or $\lambda_{c,\phi}$ increases), the probability of failure obtained through all the copulas becomes smaller. This behavior makes sense since a decrease in the C.O.V. means a decrement of the standard deviation (keeping the mean unchanged), and this leads to less heavy-tailed multivariate distributions.
- As Kendall's tau coefficient approaches to -1, the probability of failure decreases. This behavior is expected since the closer the dependence approaches perfection, the less dispersion is in the final distribution, which leads to smaller probabilities of failure.
- For the same geotechnical model and dependence value between random variables, different copulas lead to different probabilities of failure. Discrepancies among their results are more pronounced when the reliability of the model is high, i.e. for small probabilities of failure. In contrast, it is seen that probabilities of failure obtained from different copulas approach to a same value when the reliability of the model is low, i.e. for high probabilities of failure.
- Differences among probabilities of failure are more sensitive to changes in statistical parameters than in geometric parameters. That is to say, the highest discrepancies among probabilities of failure are found when the C.O.V., or the dependence parameter τ of the random variables change.
- Note from figure 7.2 that the Gaussian copula leads to the smallest probabilities of failure for all the evaluated conditions. Different studies agree that the lowest probabilities of failure, or the highest reliability indexes [83], are always obtained from the Gaussian copula for each particular geotechnical model analyzed. This fact indicates that the use of this

copula, or the commonly employed multivariate normal distribution, overestimates the reliability, and hence, it may be too unconservative.

- Among all the studied conditions and parameters, changes in Kendall's tau are the ones that lead to the highest discrepancies in the probability of failure obtained from different copulas. This fact shows the importance of correctly fitting the copulas since the probability of failure is quite sensitive to this parameter.
- Note that all the studied copulas can approach both a perfect negative dependence and independence. Thus, when τ approaches to 0, all the obtained probabilities of failure should converge to the same value, corresponding to the one obtained from the independence copula $\Pi(u,v)$. On the other hand, when τ approaches to -1, all probabilities of failure should approach the same value, given by the countermonotonicity copula $W(u,v)$. Although this behavior was not explicitly evidenced in this document, it was pointed out by other investigations, such as Tang et al. [73].
- Although both Student-t and Gaussian copulas are elliptical copulas, their probabilities of failure differ significantly, especially in high reliability levels. These discrepancies are mainly due to the tail dependence of the Student-t copula. In this way, tail dependence plays an important role when performing reliability analyses, since the computed probabilities of failure may vary in several orders of magnitude by taking it into account or not.
- In this particular infinite slope example, the MCS performs well in comparison to the direct integration method applied by Tang et al. [85]. However, the true benefits of the MCS over the direct integration method appear when the performance function is highly non-linear or when there are more than four random variables involved in the reliability analysis. In this way, the MCS is a more general methodology than the direct integration method, and its use is recommended for the computation of probabilities of failure.

Finally, it is worth mentioning that some studies such as Wu [77, 82] or Zhang et al. [83], prefer to express the reliability in terms of reliability indexes instead of probabilities of failure. Particularly, Zhang et al. [83] noted that it is preferable to employ the reliability index when performing a weighting of the results, in order to avoid disproportional effects given by the large differences between probabilities of failure. Unlike probabilities of failure of different copulas, where the discrepancies between them may be of several orders of magnitude, differences in reliability indexes are not as pronounced. Commonly, all the reliability indexes of a geotechnical model obtained through different copulas are in the same order of magnitude, which makes their discrepancies minor.

7.3 Efforts to unify results of reliability analysis of geotechnical structures when dependence is considered.

In the last section, it was demonstrated through an infinite slope example that different copulas lead to different values of probabilities of failure. Particularly, for a case of very low probabilities of failure, the results obtained from different copulas may differ by several orders of magnitude. Based on this, the question that arises is which is the true or real reliability of the geotechnical system? or which of all these probabilities of failure better represent the reliability?.

Firstly, one could think about selecting the probability of failure given by the copula that best fits data, which in turn is chosen through one of the methods of section 4.5. However, although this is a valid option, it was demonstrated in section 6 that defining the best-fit copula could be misleading, especially when no sufficient data is available. In other words, there is not a complete certainty of choosing the best-fit copula when the amount of data is small, which is quite common in geotechnical engineering practice.

As noted by Zhang et al. [83], there would not be model selection uncertainty if the probability of one model being the real one were 1.0, while the probabilities of the others were zero. This fact seldom occurs, but what is certain is that there are models more likely to be the correct ones in contrast to the others. Nonetheless, it can not be said under these circumstances that these high-likely models have the final word regarding the probability of failure since the inherent model uncertainty would be denied.

A good estimation of the uncertainty when selecting models is through the ratio $P(M_1|D)/P(M_2|D)$, where $P(M|D)$ is the probability of model M being the true one, given the data D. If this ratio is equal or larger than 20, the hypothesis of M_1 being overwhelmingly better than M_2 is correct. On the contrary, if this ratio is between 1.0 and 3.0 this hypothesis becomes only a vague evidence in favor of M_1 [see, e.g., 132].

The previous estimation was employed by Zhang et al. [83]. In this work several bivariate copula models were fitted to a shear strength dataset, and each model probability was computed based on a Bayesian approach. As expected, some copula models (with their respective marginal distributions) give a better fit to data than other copulas. However, no model overwhelmingly fits better with data in comparison to the others. In particular, the ratio between the probabilities of the most likely models is just 1.1, which indicates that model uncertainties can not be neglected.

In order to characterize the dispersion between the probabilities of failure obtained through different copulas in a family $\mathscr{C}$, the global dispersion factor r [see 31], was employed by Tang et al. [73] in an example of an infinite homogeneous slope. The formula of the *global dispersion factor* is given by:

$$r = \frac{P_{f_{max}}}{P_{f_{min}}}$$

where $P_{f_{max}} = \max_{\mathtt{C}\in\mathscr{C}} P_f(\mathtt{C})$ and $P_{f_{min}} = \min_{\mathtt{C}\in\mathscr{C}} P_f(\mathtt{C})$.

A small r indicates that the probabilities of failure vary in a narrow range, while a large r means a substantial dispersion between the probabilities of failure. As stated by Tang et al. [73], if $1.0 \leq r \leq 1.5$, the probability of failure can be evaluated quantitatively. On the other hand, if $1.5 < r \leq 10.0$, the probability of failure can be evaluated qualitatively. And finally, if $r > 10.0$, then the estimated probability of failure may exceed the real probability of failure by at least one order of magnitude, which may be too inaccurate and dangerous when defining the reliability of a geotechnical structure.

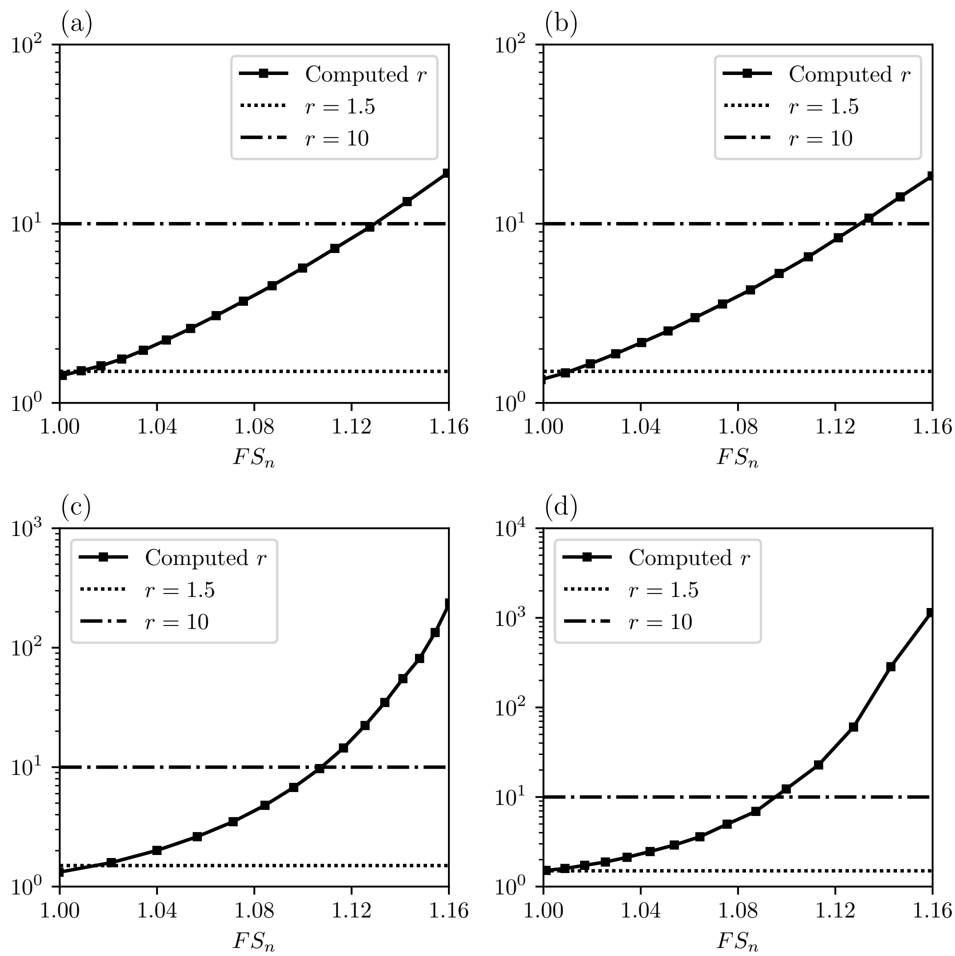


Figure 19 Global dispersion factors of the studied infinite slope in section 7.2. (a) H varying from 6.27 to 3.84 m, (b) α varying from 32° to 27.7°, (c) $\lambda_{c,\phi}$ varying from 0.67 to 1.64, and (d) $\tau_{c,\phi}$ varying from -0.26 to -0.78

Other good estimate of the dispersion in the results obtained from different copulas is the local dispersion factor r_0. The *local dispersion factor* refers to the relation between the $P_{f_{max}}$ (or $P_{f_{min}}$) and the probability obtained from certain copula $\mathtt{C}$ in $\mathscr{C}$. Its formula is expressed as follows:

$$r_0 = \max\left\{\frac{P_{f_{max}}}{P_f(\mathtt{C})}, \frac{P_f(\mathtt{C})}{P_{f_{min}}}\right\}$$

Note that the global dispersion factor quantifies the maximum dispersion in the probabilities of failure obtained from the copulas of $\mathscr{C}$, whereas the local dispersion factor quantifies the maximum dispersion concerning a reference copula in $\mathscr{C}$.

It is interesting to compute and plot the global dispersion factors from the homogeneous infinite slope example of section 7.2. This procedure was performed and figure 7.3 presents a curve for each condition studied in the last section that represents the global dispersion factor.

As exposed previously, differences among the probabilities of failure obtained through each copula (i.e., $P_f(\mathtt{C})$), are more sensitive to changes in τ or λ than to changes in geometric parameters. This behavior is evident in figure 7.3, since global dispersion factors are greater for these probabilistic parameters, and particularly, they are greater for changes in τ.

Now, it is interesting to compare, with help of the figure 7.3, the obtained global dispersion factors with the intervals defined by Tang et al. [73]. Following this guideline, it is found that for all the studied conditions the probability of failure can be evaluated quantitatively just for the smallest factor of safety, close to 1.0. For larger factors of safety (≈ 1.02 or higher) the probability of failure must be evaluated qualitatively. And, for the highest factors of safety (≈ 1.12 or higher), the estimated probability of failure may exceed the real probability by at least one order of magnitude. These comparisons show great dispersions between results obtained through different copulas.

The above justifies the need for a more consistent estimation of the probability of failure, which has to take into account the inherent uncertainty. For this purpose, several methodologies have been employed in the literature on copula theory in geotechnical engineering. Specifically, we found that three methodologies presented by Tang et al. [73] to overcome this issue, which in turn are based on the postulates of Dutfoy and Lebrun [31].

Taking into account that the three methodologies exposed by Tang et al. [73] serve to understand other methodologies and approaches that complement them, these will be briefly presented below. Subsequently, comparisons are made with the approaches and modifications presented in other studies, such as Zhang et al. [83] or Xu et al. [74]. Thus, the three methodologies of Tang et al. [73] are summarized as follows:

1. The first approach consists of, conservatively, taking the copula in $\mathscr{C}$ which leads to the highest estimation of the probability of failure. It is common to see this practice in geotechnical engineering, especially when the amount of data is scarce, and hence there is high uncertainty. However, its major disadvantage is that it could be too conservative, which increases the robustness of the designs/solutions, and hence their price significantly.
2. The second approach is based on the local dispersion factor defined by the equation 7.3. Briefly, the copula belonging to $\mathscr{C}$ that leads to the smallest local dispersion factor must be selected, and hence, its probability of failure must be adopted. This approach minimizes local dispersions in the final probability of failure and provides more reasonable estimates.
3. The third approach, and perhaps the most appealing, consists of a weighted average of all the copulas belonging to $\mathscr{C}$. With this approach, a new copula is constructed in the form of:

$$\mathtt{C}(u,v;\ \theta) = \sum_{i=1}^{n} p_i \mathtt{C}_i(u,v;\ \theta_i) \tag{35}$$

   where p_i is the probability of each copula to be the correct one, so $\sum_{i=1}^{n} p_i = 1.0$, and $\mathtt{C}_i$ is the copula i-th in $\mathscr{C}$, with its respective fitting parameters represented by the vector θ_i.
   This approach has the advantage of being straightforward but, at the same time, robust. It allows taking into account every copula of $\mathscr{C}$ and to make estimates of, e.g., the probability of failure of the model. In the final estimates, each candidate copula has a certain incidence that may be greater or less, in concordance with the respective probability of being the correct copula.

Regarding the probability of each copula being the correct one (p_i in the equation (35)), there have been exposed several methodologies in the literature to address the computation of this value. For example, the procedure presented by Tang et al. [73] for this purpose is based on a bootstrap approach, quite similar to what was done in section 6. In this sense, Tang et al. [73] employed the shear strength dataset, reported by Tang et al. [85], to conduct a bootstrap procedure. A total of ten thousand bootstrap samples ($N_s = 10000$) were recreated by Tang et al. [73], and for each one, the AIC value of the different copulas was computed. Finally, the number of times that each copula is defined as the best fit copula is divided by the number of bootstrap samples (N_z), which leads to the probability of each copula being the correct one (p_i).

Another way to compute p_i values from equation (35) was presented by Zhang et al. [83], which in turn was applied over a shear strength dataset to demonstrate the procedure. In this work, the Bayes' theorem, based on a set of observed data (D), is employed to calculate the posterior probability of copula model M_i being the correct one, as follows:

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{\sum_{j=1}^{r} P(D|M_j)P(M_j)} \tag{36}$$

where $P(M_i|D)$ is posterior probability of M_i being the true model, $P(D|M_i)$ is the probability of observe D given that the model is M_i, and $P(M_i)$ is the prior probability that the model M_i is the true one. If no prior probability is specified, then it could be assumed that $P(M_i) = \frac{1}{r}$, where r is the number of candidate models in $\mathscr{C}$.

Note that $P(M_i|D)$ in equation (36) is the same p_i of equation (35). Additionally, after some approximations, and based on Kass and Raftery [132], equation (36) can be rewritten as follows [see 83]:

$$P(M_i|D) \approx \frac{\exp\left[-\Delta_i(BIC)/2\right]}{\sum_{j=1}^{r} \exp\left[-\Delta_j(BIC)/2\right]}$$

in which

$$\Delta_i(BIC) = BIC_i - \min_{j=1,2\ldots,r} \left\{BIC_j\right\}$$

where BIC is the Bayesian Information Criterion exposed in section 4.5.2. This approximation makes it easier to compute the probability of each model (p_i), which makes this method as straightforward and easily-applicable, just like the bootstrap approach employed by Tang et al. [73].

Another approach to compute p_i that is worth mentioning is the entropy weights [133]. This approach was applied in geotechnics by Xu et al. [74] to weigh the results of different copulas, applying equation (35), when calculating the probability of failure of a slope with a rotational failure mechanism. Note that all the aforementioned weighting methods vary only in the manner of computing p_i. Thus, the principle by itself is equal, since the calculated p_i are replaced in the equation (35) to weight the results obtained through all the copulas.

In this way, equation (35) has been applied by several studies to weight results obtained from different copulas, specifically when these are employed to conduct geotechnical reliability analysis. The main difference of some studies when applying equation (35) is the way in which these compute the probability of each model to be the correct one p_i.

In conclusion, from our point of view the second and third methods, presented by Tang et al. [73], are the most adequate to take into account the model uncertainty in the estimates of the final probability of failure. Between these two methods, the one based on the weighted average may be more appealing than the one based on the local dispersion factor. The first method presented by Tang et al. [73], although it is a common practice in geotechnical engineering, it is also too conservative and must be avoided whenever it is possible.

Although the aforementioned approaches take into account model uncertainties, and certainly these compute more robust estimates of the probability of failure, they are not perfect methodologies. In fact, more research is needed in this regard, since model uncertainty, especially when the amount of data is small, is still a challenging problem.

8 Conclusions

Throughout this document, copula theory and its applicability for the modeling of dependence in geotechnical engineering have been exposed. For this purpose, copulas and basic concepts of dependence in geotechnical engineering were introduced from a theoretical and a practical perspective. Concepts about uncertainty in the construction and definition of copulas, as well as their impact on models, were also included in this state-of-the-art. Finally, the role that copulas have had in geotechnical reliability assessments was also evaluated. Based on this work, the following conclusions can be drawn:

1. Dependence plays a key role in geotechnical engineering and, therefore, it should not be taken lightly, much less overlooked. Dependence on geotechnical parameters impacts, directly or indirectly, the constructed models in a significant way.
2. The practice of geotechnics is full of equations that correlate different parameters and properties of soils and rocks. However, using these equations leads to many uncertainties in the estimates sought and, therefore, in the models built.
3. It is necessary to migrate, in geotechnical engineering, from the concept of correlation to the concept of dependence, which is much broader and, in fact, encompasses correlations. Starting to consider dependence will allow a much better understanding of how geotechnical parameters are related.
4. Copula theory is especially suited for modeling dependence. By means of this theory, a multivariate CDF can be separated into marginal CDFs and a function that groups them together, known as the copula function. This fact is quite useful since it opens up a fairly wide range of possibilities when it comes to modeling multivariate distribution functions and studying the dependence structures between random variables.
5. It is quite common in geotechnics to find that soils and rocks parameters, and in general all random variables, are related in pairs. In many cases, it is enough to use bivariate copulas to model the pairs of random variables. If it is necessary to build models of more variables, they could also be constructed with the conventional copula theory, although this has certain disadvantages related to the assignment of the same dependence structure for all the variables, and that the range of available copulas is reduced. In this case, it would be much more convenient to make use of vine copula theory.
6. There is no ideal copula to model dependence on geotechnics. Actually, there are multiple copulas in the literature, of which a reduced set must be selected based on the dependence characteristics that we must model. In any case, each copula must be adjusted to the information and the most suitable copula must be chosen through a goodness-of-fit test. However, there will always be some uncertainties in the construction and selection of a copula function, especially when the information is limited.
7. There will always be uncertainties in the construction and selection of copulas for modeling dependencies in geotechnical engineering. These uncertainties will be greater as the quality and quantity of the information decreases. Thus, the uncertainties will be transmitted to the created models and the results obtained from them. Therefore, it is necessary to build and select copulas responsibly, in order to obtain models that truthfully represent reality.
8. When conducting geotechnical reliability analysis, different copulas will lead to different results in the computed probability of failure. As it was demonstrated, probabilities of failure obtained from different copulas can vary by several orders of magnitude, especially when P_f is small. This is one more example of the impact of epistemic uncertainty, and it highlights the importance of selecting copulas responsibly. More research should be carried out on the correct construction and selection of copulas to model geotechnical variables.
9. Gaussian dependence structure, commonly assumed in geotechnical probability models without any validation, is not always the most suitable copula function. In reality, this dependence structure is only one of many that can be employed in light of copula theory. In the case of reliability analyses, it was shown that Gaussian copula may lead to overestimations of the probability of failure, which becomes too conservative, especially when the probability of failure is small.
10. Monte Carlo simulation is one of the best alternatives to carry out reliability analyses in geotechnics within the framework of copula theory. It is important to be careful with the number of simulations employed in this technique since an insufficient number would lead to incorrect reliability estimates.
11. There are simulation techniques, much more advanced than Monte Carlo simulation, which could be used together with copula theory. In particular, subset simulation is a modern technique that allows to calculate small probabilities of failure in a much more efficient way than Monte Carlo simulation. Some studies have already integrated subset simulation with copulas.

All data and codes used in this manuscript are publicly available on GitHub at https://github.com/juanjosesg/geo-copulas.git.

Open issues and directions for future research

Although the uses of copula theory in geotechnics have been varied, there is still much to explore. As it was previously mentioned, geotechnics is an area where correlations abound and where independence or Gaussian dependence are assumed without validation. For this reason, copula theory would find a wide application in geotechnics, as it has already been demonstrated with the applications that have been evaluated so far.

There are multiple uses that copula theory may have in geotechnical engineering, and those studied here are just some of them. In this way, the concepts presented in this document can be applied in many other contexts of geotechnics. For example, copulas can be used in the study of dependence between parameters of different constitutive models and failure criteria, it could be employed in many of the already existing correlations (e.g., those of the SPT with different geotechnical parameters), and it might even be employed in the study of the relationships between rains, earthquakes, and landslides, among many other uses. The applications of copulas presented in this document are just a hint of the great potential that copulas have in geotechnics.

Furthermore, vine copula theory has great potential that should be explored in future works. This theory allows to generate high-dimensional models through two-dimensional copulas. By using this theory, it is possible to assign particular dependence structures for each pair of random variables and, as bivariate copulas are employed, construct flexible high-dimensional dependence models.

Finally, selection of the best-fit copula is still an open issue in the state-of-the-art. There are some developments in this regard, but the final choice is still highly dependent on the quality and quantity of available information.

References

[1] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? Does it matter? *Structural safety*, 31(2):105–112, 2009.

[2] Michael Beer, Yi Zhang, Ser Tong Quek, and Kok Kwang Phoon. Reliability analysis with scarce information: Comparing alternative approaches in a geotechnical engineering context. *Structural Safety*, 41:1–10, 2013.

[3] A. Nataf. Détermination des distributions de probabilités dont les marges sont données. *C. R. Acad Sci*, 225:42–43, 1962.

[4] Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

[5] Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton University Press, 2015.

[6] Umberto Cherubini, Elisa Luciano, and Walter Vecchiato. *Copula methods in finance*. John Wiley & Sons, 2004.

[7] Lu Chen, Vijay P Singh, and Shenglian Guo. Measure of correlation between river flows using the copula-entropy method. *Journal of Hydrologic Engineering*, 18(12): 1591–1606, 2013.

[8] Subimal Ghosh. Modelling bivariate rainfall distribution and generating bivariate correlated rainfall data in neighbouring meteorological subdivisions using copula. *Hydrological Processes*, 24(24):3558–3567, 2010.

[9] LSVP Zhang and VP Singh. Bivariate flood frequency analysis using the copula method. *Journal of Hydrologic Engineering*, 11(2):150–164, 2006.

[10] K Goda. Statistical modeling of joint probability distribution using copula: application to peak and permanent displacement seismic demands. *Structural Safety*, 32(2): 112–123, 2010.

[11] Roger B Nelsen. *An introduction to copulas*. Springer, 2007.

[12] Paul Embrechts, Filip Lindskog, and Alexander McNeil. Modelling dependence with copulas and applications to risk management. *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich*, 14, 2001.

[13] Paul Embrechts, Alexander McNeil, and Daniel Straumann. Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, 1:176–223, 2002.

[14] Abe Sklar. Random variables, distribution functions, and copulas: A personal look backward and forward. *Lecture Notes-Monograph Series*, 28:1–14, 1996. ISSN 07492170. URL http://www.jstor.org/stable/4355880.

[15] Ove Ditlevsen and Henrik O Madsen. *Structural reliability methods*, volume 178. Wiley New York, 1996.

[16] Tim Bedford and Roger M Cooke. Vines: A new graphical model for dependent random variables. *Annals of Statistics*, pages 1031–1068, 2002.

[17] Kjersti Aas, Claudia Czado, Arnoldo Frigessi, and Henrik Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2): 182–198, 2009.

[18] Claudia Czado. Analyzing dependent data with vine copulas. *Lecture Notes in Statistics, Springer*, 2019.

[19] Tim Bedford and Roger M Cooke. Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32(1):245–268, 2001.

[20] Alfredo Hua-Sing Ang and Wilson H Tang. *Probability concepts in engineering planning and design: Emphasis on application to civil and environmental engineering*. Wiley, 2007.

[21] Jay Ameratunga, Nagaratnam Sivakugan, and Braja M Das. *Correlations of soil and rock properties in geotechnical engineering*. Springer, 2016.

[22] Armen Der Kiureghian and Pei-Ling Liu. Structural reliability under incomplete probability information. *Journal of Engineering Mechanics*, 112(1):85–104, 1986.

[23] Pei-Ling Liu and Armen Der Kiureghian. Multivariate distribution models with prescribed marginals and covariances. *Probabilistic Engineering Mechanics*, 1(2): 105–112, 1986.

[24] Hong Shuang Li, Zhen Zhou Lü, and Xiu Kai Yuan. Nataf transformation based point estimate method. *Chinese Science Bulletin*, 53(17):2586, 2008.

[25] Regis Lebrun and Anne Dutfoy. An innovating analysis of the Nataf transformation from the copula viewpoint. *Probabilistic Engineering Mechanics*, 24(3):312–320, 2009.

[26] Régis Lebrun and Anne Dutfoy. A generalization of the Nataf transformation to distributions with elliptical copula. *Probabilistic Engineering Mechanics*, 24(2): 172–178, 2009.

[27] Douglas C Montgomery and George C Runger. *Applied statistics and probability for engineers*. John Wiley & Sons, 2010.

[28] Nathabandu T Kottegoda and Renzo Rosso. *Applied statistics for civil and environmental engineers*. Blackwell Malden, MA, 2008.

[29] Kumar Jogdeo. Concepts of dependence. *Encyclopedia of statistical sciences*, 1:324–334, 1982.

[30] Samuel Kotz and Dominique Drouet. *Correlation and dependence*. World Scientific, 2001.

[31] A Dutfoy and R Lebrun. Practical approach to dependence modelling using copulas. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 223(4):347–361, 2009.

[32] Harry Joe. *Multivariate models and multivariate dependence concepts*. CRC Press, 1997.

[33] Edward W. Frees and Emiliano A. Valdez. Understanding relationships using copulas. *North American Actuarial Journal*, 2(1):1–25, 1998.

[34] Christian Genest and Jock MacKay. The joy of copulas: bivariate distributions with uniform marginals. *The American Statistician*, 40(4):280–283, 1986.

[35] Kok-Kwang Phoon and Jianye Ching. *Risk and reliability in geotechnical engineering*. CRC Press, 2014.

[36] Lan Zhang and Vijay P. Singh. *Copulas and their applications in water resources engineering*. Cambridge University Press, 2019.

[37] Dian Qing Li, Xiao Song Tang, Kok Kwang Phoon, Yi Feng Chen, and Chuang Bing Zhou. Bivariate simulation using copula and its application to probabilistic pile settlement analysis. *International Journal for Numerical and Analytical Methods in Geomechanics*, 37(6): 597–617, 2013. ISSN 03639061.

[38] Christian Genest and Anne-Catherine Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12 (4):347–368, 2007.

[39] Valdo Durrleman, Ashkan Nikeghbali, and Thierry Roncalli. Which copula is the right one? *SSRN Electronic Journal*, 2000. doi: 10.2139/ssrn.1032545.

[40] Berthold Schweizer and Edward F. Wolff. On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9(4):879–885, 1981.

[41] Henrik Hult and Filip Lindskog. Multivariate extremes, aggregation and dependence in elliptical distributions. *Advances in Applied Probability*, 34(3):587–608, 2002.

[42] Harry Joe and James Jianmeng Xu. The estimation method of inference functions for margins for multivariate models. Technical report, University of British Columbia, 1996. URL https://open.library.ubc.ca/collections/52383/items/1.0225985.

[43] Christian Genest, Bruno Rémillard, and David Beaudoin. Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44(2): 199–213, 2009.

[44] Yannick Malevergne and Didier Sornette. Testing the Gaussian copula hypothesis for financial assets dependences. *Quantitative Finance*, 3(4):231–250, 2003.

[45] Christian Genest and Louis-Paul Rivest. Statistical inference procedures for bivariate archimedean copulas. *Journal of the American Statistical Association*, 88(423): 1034–1043, 1993.

[46] Philippe Barbe, Christian Genest, Kilani Ghoudi, and Bruno Rémillard. On Kendall's process. *Journal of Multivariate Analysis*, 58(2):197–229, 1996.

[47] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on Automatic Control*, 19(6): 716–723, 1974.

[48] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[49] Kenneth P. Burnham and David R. Anderson. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304, 2004.

[50] Scott I Vrieze. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, 17(2):228, 2012.

[51] Kenneth P. Burnham and David R. Anderson. A practical information-theoretic approach. *Model selection and multimodel inference, 2nd ed. Springer, New York*, 2002.

[52] T. William Lambe and Robert V. Whitman. *Soil mechanics*, volume 10. John Wiley & Sons, 1991.

[53] James Kenneth Mitchell and Kenichi Soga. *Fundamentals of Soil behavior*, volume 3. John Wiley & Sons New York, 2005.

[54] B.K. Low. Reliability analysis of rock slopes involving correlated nonnormals. *International Journal of Rock Mechanics and Mining Sciences*, 44(6):922–935, 2007.

[55] Dianqing Li, Yifeng Chen, Wenbo Lu, and Chuangbing Zhou. Stochastic response surface method for reliability analysis of rock slopes involving correlated non-normal variables. *Computers and Geotechnics*, 38(1):58–68, 2011.

[56] Xiao-Song Tang, Dian-Qing Li, Yi-Feng Chen, Chuang-Bing Zhou, and Li-Min Zhang. Improved knowledge-based clustered partitioning approach and its application to slope reliability analysis. *Computers and Geotechnics*, 45:34–43, 2012.

[57] Peter Lumb. Safety factors and the probability distribution of soil strength. *Canadian Geotechnical Journal*, 7(3): 225–242, 1970.

[58] Thomas Francis Wolff. *Analysis and design of embankment dam slopes: a probabilistic approach*. University Microfilms, 1985.

[59] Minoru Matsuo and Katsuhiko Kuroda. Probabilistic approach to design of embankments. *Soils and Foundations*, 14(2):1–17, 1974.

[60] C. Cherubini. Reliability evaluation of shallow foundation bearing capacity on c - ϕ soils. *Canadian Geotechnical Journal*, 37(1):264–269, 2000.

[61] Gordon A. Fenton and D.V. Griffiths. Bearing-capacity prediction of spatially random c-ϕ soils. *Canadian Geotechnical Journal*, 40(1):54–65, 2003.

[62] Wolfgang Fellin and Michael Oberguggenberger. Robust assessment of shear parameters from direct shear tests. *International Journal of Reliability and Safety*, 6(1-3): 49–64, 2012.

[63] Eduardo E. Alonso. Risk analysis of slopes and its application to slopes in canadian sensitive clays. *Geotechnique*, 26(3):453–472, 1976.

[64] D.C. Tobutt and E.A. Richards. The reliability of earth slopes. *International Journal for Numerical and Analytical Methods in Geomechanics*, 3(4):323–354, 1979.

[65] K.S. Li and P. Lumb. Probabilistic design of slopes. *Canadian Geotechnical Journal*, 24(4):520–535, 1987.

[66] Ya-Fen Lee and Yun-Yao Chi. Rainfall-induced landslide risk at Lushan, Taiwan. *Engineering Geology*, 123(1-2): 113–121, 2011.

[67] V.U. Nguyen and R.N. Chowdhury. Probabilistic study of spoil pile stability in strip coal mines—two techniques compared. In *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts*, volume 21, pages 303–312. Elsevier, 1984.

[68] Heidy Sánchez-Lizarraga and Carlo G. Lai. Effects of spatial variability of soil properties on the seismic response of an embankment dam. *Soil Dynamics and Earthquake Engineering*, 64:113–128, 2014.

[69] Chris Parker, Andrew Simon, and Colin R. Thorne. The effects of variability in bank material properties on riverbank stability: Goodwin Creek, Mississippi. *Geomorphology*, 101(4):533–543, 2008.

[70] Y. Hata, K. Ichii, T. Tsuchida, S. Kano, and N. Yamashita. A practical method for identifying parameters in the seismic design of embankments. *Georisk*, 2(1):28–40, 2008.

[71] G.B. Baecher and J.T. Christian. *Reliability and Statistics in Geotechnical Engineering*. John Wiley & Sons, 2005. ISBN 9780471498339.

[72] Abdul-Hamid Soubra and Nut Mao. Probabilistic analysis of obliquely loaded strip foundations. *Soils and Foundations*, 52(3):524–538, 2012.

[73] Xiao Song Tang, Dian Qing Li, Chuang Bing Zhou, and Kok Kwang Phoon. Copula-based approaches for evaluating slope reliability under incomplete probability information. *Structural Safety*, 52:90–99, 2015. ISSN 01674730. doi: 10.1016/j.strusafe.2014.09.007.

[74] Xiaoliang Xu, Jianlin Li, Jiawei Gong, Huafeng Deng, and Liangpeng Wan. Copula-based slope reliability analysis using the failure domain defined by the g-line. *Mathematical Problems in Engineering*, 2016, 2016. ISSN 15635147. doi: 10.1155/2016/6141838.

[75] Xiao Song Tang, Dian Qing Li, Zi Jun Cao, and Kok Kwang Phoon. Impact of sample size on geotechnical probabilistic model identification. *Computers and Geotechnics*, 87:229–240, 2017. ISSN 18737633. doi: 10.1016/j.compgeo.2017.02.019.

[76] Xing Zheng Wu. Probabilistic slope stability analysis by a copula-based sampling method. *Computational Geosciences*, 17(5):739–755, 2013. ISSN 14200597. doi: 10.1007/s10596-013-9353-3.

[77] Xing Zheng Wu. Trivariate analysis of soil ranking-correlated characteristics and its application to probabilistic stability assessments in geotechnical engineering problems. *Soils and Foundations*, 53(4):540–556, 2013. doi: 10.1016/j.sandf.2013.06.006.

[78] Songsak Sriboonchitta and Vladik Kreinovich. Why are FGM copulas successful? a simple explanation. *Advances in Fuzzy Systems*, 2018, 2018.

[79] Lei Zhang, XS Tang, and DQ Li. Bivariate distribution model of soil shear strength parameter using copula. *Journal of Civil Engineering and Management*, 30(2):11–17, 2013.

[80] Dian Qing Li, Lei Zhang, Xiao Song Tang, Wei Zhou, Jin Hui Li, Chuang Bing Zhou, and Kok Kwang Phoon. Bivariate distribution of shear strength parameters using copulas and its impact on geotechnical system reliability. *Computers and Geotechnics*, 68:184–195, 2015. ISSN 18737633. doi: 10.1016/j.compgeo.2015.04.002.

[81] Q.B. Yu. Slope reliability of embankment dam and its application to engineering practice. Master's thesis, Hohai University, Nanjing, China, 2006.

[82] Xing Zheng Wu. Modelling dependence structures of soil shear strength data with bivariate copulas and applications to geotechnical reliability analysis. *Soils and Foundations*, 55(5):1243–1258, 2015. ISSN 00380806. doi: 10.1016/j.sandf.2015.09.023.

[83] J. Zhang, H. W. Huang, C. H. Juang, and W. W. Su. Geotechnical reliability analysis with limited data: consideration of model selection uncertainty. *Engineering Geology*, 181:27–37, 2014. ISSN 00137952. doi: 10.1016/j.enggeo.2014.08.002.

[84] Da Huang, Chao Yang, Bin Zeng, and Guoyang Fu. A copula-based method for estimating shear strength parameters of rock mass. *Mathematical Problems in Engineering*, 2014, 2014. ISSN 15635147. doi: 10.1155/2014/693062.

[85] Xiao Song Tang, Dian Qing Li, Guan Rong, Kok Kwang Phoon, and Chuang Bing Zhou. Impact of copula selection on geotechnical reliability under incomplete probability information. *Computers and Geotechnics*, 49:264–278, 2013. ISSN 0266352X. doi: 10.1016/j.compgeo.2012.12.002.

[86] Xu Chao and YANG Lin-de. Test of goodness of fit of random variables and Bayesian estimation of distribution parameters. *Journal of Tongji University*, 26(3):340–344, 1998.

[87] M Dithinde, K.K. Phoon, M. De Wet, and J.V. Retief. Characterization of model uncertainty in the static pile design formula. *Journal of Geotechnical and Geoenvironmental Engineering*, 137(1):70–85, 2011.

[88] Jonathan C. Huffman, Andrew W. Strahler, and Armin W. Stuedlein. Reliability-based serviceability limit state design for immediate settlement of spread footings on clay. *Soils and Foundations*, 55(4):798–812, 2015.

[89] Marco Uzielli and Paul W. Mayne. Serviceability limit state CPT-based design for vertically loaded shallow footings on sand. *Geomechanics and Geoengineering*, 6(2): 91–107, 2011. ISSN 17486025. doi: 10.1080/17486025.2010.531146.

[90] Marco Uzielli and Paul W. Mayne. Load-displacement uncertainty of vertically loaded shallow footings on sands and effects on probabilistic settlement estimation. *Georisk*, 6(1):50–69, 2012. ISSN 17499518. doi: 10.1080/17499518.2011.626333.

[91] Paul W Mayne and Harry G Poulos. Approximate displacement influence factors for elastic shallow foundations. *Journal of Geotechnical and Geoenvironmental Engineering*, 125(6):453–460, 1999.

[92] Jean-Louis Briaud. Spread footings in sand: load settlement curve approach. *Journal of Geotechnical and Geoenvironmental Engineering*, 133(8):905–920, 2007.

[93] Kok-Kwang Phoon, Fred H. Kulhawy, and Mircea D. Grigoriu. Multiple resistance factor design for shallow transmission line structure foundations. *Journal of Geotechnical and Geoenvironmental Engineering*, 129(9): 807–818, 2003.

[94] K.K. Phoon, J.R. Chen, and F.H. Kulhawy. Characterization of model uncertainties for augured cast-in-place piles under axial compression. In *Foundation Analysis and Design: Innovative Methods*, pages 82–89. 2006.

[95] Dianqing Li, Xiaosong Tang, Chuangbing Zhou, and Kok Kwang Phoon. Uncertainty analysis of correlated non-normal geotechnical parameters using Gaussian copula. *Science China Technological Sciences*, 55(11): 3081–3089, 2012. ISSN 16747321. doi: 10.1007/s11431-012-4937-z.

[96] P.K. Robertson and K.L. Cabal. *Guide to cone penetration testing for geotechnical engineering*. Gregg Drilling & Testing, Inc., 2015.

[97] Jonathan C Huffman and Armin W Stuedlein. Reliability-based serviceability limit state design of spread footings on aggregate pier reinforced clay. *Journal of Geotechnical and Geoenvironmental Engineering*, 140(10):04014055, 2014.

[98] Fung Kee Chin. Estimation of the ultimate load of piles from tests not carried to failure. In *Proceedings, 2nd Southeast Asian Conference on Soil Engineering*, Singapore, 1970.

[99] M.T. Davison. High-capacity piles. In *Proceedings, Lecture Series, Innovations in Foundation Construction*, Chicago, 1972. ASCE, Illinois Section.

[100] K.K. Phoon, J.-R. Chen, and F.H. Kulhawy. Probabilistic hyperbolic models for foundation uplift movements. In *Probabilistic Applications in Geotechnical Engineering*, pages 1–12. 2007.

[101] Kok-Kwang Phoon. *Reliability-based design in geotechnical engineering: computations and applications*. CRC Press, 2008.

[102] Dian-Qing Li, Xiao-Song Tang, Chuang-Bing Zhou, and Kok-Kwang Phoon. Characterization of uncertainty in probabilistic model using bootstrap method and its application to reliability of piles. *Applied Mathematical Modelling*, 39(17):5310–5326, 2015.

[103] Steven Lawrence Kramer. *Geotechnical earthquake engineering*. Pearson Education India, 1996.

[104] C. Allin Cornell. Engineering seismic risk analysis. *Bulletin of the Seismological Society of America*, 58(5): 1583–1606, 1968.

[105] Robin K. McGuire. *Seismic hazard and risk analysis*. Earthquake Engineering Research Institute, 2004.

[106] K. Goda and G.M. Atkinson. Interperiod dependence of ground-motion prediction equations: A copula perspective. *Bulletin of the Seismological Society of America*, 99: 922–927, 2009.

[107] Yun Xu, Xiao Song Tang, J. P. Wang, and H. Kuo-Chen. Copula-based joint probability function for PGA and CAV: a case study from Taiwan. *Earthquake Engineering and Structural Dynamics*, 45(13):2123–2136, 2016. ISSN 10969845. doi: 10.1002/eqe.2748.

[108] Jack W Baker and C Allin Cornell. Correlation of response spectral values for multicomponent ground motions. *Bulletin of the Seismological Society of America*, 96 (1):215–227, 2006.

[109] Jack W. Baker. Correlation of ground motion intensity parameters used for predicting structural and geotechnical response. In *Tenth International Conference on Application of Statistics and Probability in Civil Engineering*, volume 8. Citeseer, 2007.

[110] Jack W Baker and Nirmal Jayaram. Correlation of spectral acceleration values from NGA ground motion models. *Earthquake Spectra*, 24(1):299–317, 2008.

[111] K Goda and Han-Ping Hong. Spatial correlation of peak ground motions and response spectra. *Bulletin of the Seismological Society of America*, 98(1):354–365, 2008.

[112] Katsuichiro Goda and Gail M Atkinson. Probabilistic characterization of spatially correlated response spectra for earthquakes in Japan. *Bulletin of the Seismological Society of America*, 99(5):3003–3020, 2009.

[113] Yin Cheng, Jun Du, and Hao Ji. Multivariate joint probability function of earthquake ground motion prediction

equations based on vine copula approach. *Mathematical Problems in Engineering*, 2020, 2020.

[114] Honggang Zhu, LM Zhang, Te Xiao, and XY Li. Generation of multivariate cross-correlated geotechnical random fields. *Computers and Geotechnics*, 86:95–107, 2017.

[115] Erik Vanmarcke. *Random fields: analysis and synthesis*. World scientific, 2010.

[116] Mao-Xin Wang, Xiao-Song Tang, Dian-Qing Li, and Xiao-Hui Qi. Subset simulation for efficient slope reliability analysis involving copula-based cross-correlated random fields. *Computers and Geotechnics*, 118:103326, 2020.

[117] Siu-Kui Au and James L Beck. Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engineering Mechanics*, 16(4): 263–277, 2001.

[118] Fan Wang and Heng Li. The role of copulas in random fields: Characterization and application. *Structural Safety*, 75:75–88, 2018.

[119] B. P. Marchant, N. P.A. Saby, C. C. Jolivet, D. Arrouays, and R. M. Lark. Spatial prediction of soil properties with copulas. *Geoderma*, 162(3-4):327–334, 2011. ISSN 00167061. doi: 10.1016/j.geoderma.2011.03.005.

[120] Hannes Kazianka and Jürgen Pilz. Copula-based geostatistical modeling of continuous and discrete data including covariates. *Stochastic environmental research and risk assessment*, 24(5):661–673, 2010.

[121] Hannes Kazianka and Jürgen Pilz. Bayesian spatial modeling and interpolation using copulas. *Computers & Geosciences*, 37(3):310–319, 2011.

[122] Manouchehr Motamedi and Robert Y. Liang. Probabilistic landslide hazard assessment using copula modeling technique. *Landslides*, 11(4):565–573, 2014. ISSN 16125118. doi: 10.1007/s10346-013-0399-z.

[123] J. P. Wang, Xiao Song Tang, Yih Min Wu, and Dian Qing Li. Copula-based earthquake early warning decision-making strategy. *Soil Dynamics and Earthquake Engineering*, 115:324–330, 2018. ISSN 02677261. doi: 10.1016/j.soildyn.2018.08.031.

[124] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

[125] Thomas Most and Tina Knabe. Reliability analysis of the bearing failure problem considering uncertain stochastic parameters. *Computers and Geotechnics*, 37(3):299–310, 2010.

[126] Zhe Luo, Sez Atamturktur, and C Hsein Juang. Bootstrapping for characterizing the effect of uncertainty in sample statistics for braced excavations. *Journal of Geotechnical and Geoenvironmental Engineering*, 139(1):13–23, 2013.

[127] Zhao-Xia Xu and Xiao-Ping Zhou. Three-dimensional reliability analysis of seismic slopes using the copula-based sampling method. *Engineering Geology*, 242: 81–91, 2018.

[128] Alfredo H-S Ang and Wilson H Tang. Probability concepts in engineering planning and design, vol. 2: Decision, risk, and reliability. *John Wiley & Sons, Inc.*, 1984.

[129] G.I. Schuëller, H.J. Pradlwarter, and Phaedon-Stelios Koutsourelakis. A critical appraisal of reliability estimation procedures for high dimensions. *Probabilistic Engineering Mechanics*, 19(4):463–474, 2004.

[130] Robert E. Melchers and André T. Beck. *Structural reliability analysis and prediction*. John Wiley & Sons, 2018.

[131] Diego A. Alvarez, Felipe Uribe, and Jorge E. Hurtado. Estimation of the lower and upper bounds on the probability of failure using subset simulation and random set theory. *Mechanical Systems and Signal Processing*, 100:782–801, 2018.

[132] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430): 773–795, 1995.

[133] Zhi-Hong Zou, Yun Yi, and Jing-Nan Sun. Entropy method for determination of weight of evaluating indicators in fuzzy synthetic evaluation for water quality assessment. *Journal of Environmental Sciences*, 18(5): 1020–1023, 2006.