Università degli Studi di Trieste

Dipartimento di Matematica Informatica e Geoscienze

Corso di Studi in Data Science and Scientific Computing

Tesi di Laurea Magistrale

# A COPULA APPROACH TO MODELING ENVIRONMENTAL EXTREME EVENTS

Laureando:
Davide Sancin

Relatore:
Prof.ssa Roberta Pappadà

ANNO ACCADEMICO 2023-2024

# Contents

**Abstract** **1**

**Sommario** **2**

**1 Introduction** **3**

**2 Copulas** **6**
2.1 Basic definitions and properties . . . . . 6
2.1.1 Mathematical definition . . . . . 6
2.1.2 The Fréchet–Hoeffding Bounds . . . . . 9
2.1.3 Sklar's theorem . . . . . 10
2.1.4 The invariance principle . . . . . 12
2.1.5 Survival copulas and symmetries . . . . . 13
2.2 Measures of association . . . . . 14
2.2.1 Rank correlation measures . . . . . 15
2.2.2 Tail dependence coefficients . . . . . 17
2.3 Copula examples . . . . . 18
2.3.1 Elliptical copulas . . . . . 18
2.3.2 Archimedean copulas . . . . . 20
2.3.3 Extreme value copulas . . . . . 22
2.4 Estimation and goodness of fit . . . . . 22
2.4.1 Estimation . . . . . 22
2.4.2 Goodness of fit . . . . . 25

**3 Copula methods for time series** **27**
3.1 Conditional copulas . . . . . 27
3.2 Marginals modeling . . . . . 28
3.3 Clustering . . . . . 30
3.3.1 Hierarchical clustering . . . . . 30
3.3.2 Copula-based hierarchical clustering . . . . . 31

**4 Data analysis** **33**
4.1 Data exploration . . . . . 34
4.2 Marginal modeling . . . . . 35
4.3 Copula fit . . . . . 38
4.3.1 Graphic exploration . . . . . 38
4.3.2 Statistical tests . . . . . 41
4.3.3 Clustering . . . . . 42
4.3.4 Estimation and Goodness of fit . . . . . 44
4.4 Findings . . . . . 46

**5 Conclusions** **48**

**Bibliography** **49**

# Abstract

In this thesis an overview of copulas as well as a possible application to the environmental field are presented. After a brief introductory chapter, in the second chapter copulas and their main properties are mathematically defined. Then, after presenting the main copula families, some methods to estimate a copula model and some statistical tests to choose the most appropriate copula family are discussed.
In the third chapter copula methods to treat multivariate time series are introduced and a method to cluster time series based on measures of association that depends only on the copula is discussed.
In the final chapter an application of the copula-based hierarichical clusters to time series of monthly rainfall maxima collected at various weather stations of the Friuli-Venezia Giulia region is presented.

# Sommario

In questa tesi viene presentata una panoramica delle copule e una loro possibile applicazione in campo ambientale. Dopo un breve capitolo introduttivo, nel secondo capitolo vengono definite matematicamente le copule e le loro proprietà più importanti. Poi, dopo aver presentato le principali famiglie di copule, vengono discussi alcuni metodi per stimare una copula e i test statistici per scegliere la famiglia di copule più appropriata.
Nel terzo capitolo vengono introdotti metodi basati sulle copule per trattare serie temporali multivariate e viene discusso un metodo per raggruppare serie temporali in base a misure di associazione che dipendono solo dalle copule.
Nel capitolo finale viene presentata un'applicazione del metodo di raggruppamento basato sulle copule su serie temporali dei massimi mensili di precipitazioni giornaliere raccolte in varie stazioni meteorologiche della regione Friuli Venezia-Giulia.

# Chapter 1

# Introduction

Copulas were first introduced by Sklar [1] in 1959, a more modern approach can be found in the books by Joe [2] and by Nelsen [3]. Copulas are used to study the dependencies between random variable, and they are particularly useful since they allow to study marginal distributions separately from their joint dependence structures. The main fields of applications are the financial and the environmental one, but they are also used in many other fields.
In many applications what is of interest is the behaviour of the distributions in the tails; it is possible to define a tail dependence coefficient, which describe the dependence in the joint tails of bivariate distributions, that depends exclusively on the underlying copula [4], essentially the tail dependence coefficient is used to model the co-occurrences of extreme events.
Since the introduction of the concept of the conditional copulas by Patton [5] in 2006, copula based methods to treat multivariate time series have gained increasingly more popularity. This is due to the fact that with the use of copulas it is possible to catch non linear and asymmetric dependencies between the components of a multivariate time series.
In particular in copula based methods for multivariate time series it is assumed that every component can be modeled as an univariate time series via classical methods such as ARIMA models and ARMA-GARCH models, and that their innovations are jointly coupled through a copula [6].
A further use of copulas in the time series context is for the clustering of time series.
The clustering of time series can be useful to find sub-groups of common behaviour as well as a way to perform a dimensionality reduction.
In hierarchical clustering the definition of a dissimilarity is required, but identifying a proper distance for time series clustering can be challenging. It

is possible to define dissimilarities for time series based on copula measures of association. These clustering methods provide all the advantages of the copulas, they should be able to catch more complex dependencies and they are rank invariant. De Luca and Zuccolotto[7] in 2011 proposed the use of the tail dependence coefficient for time series clustering. Tail dependence coefficient as a dissimilarity measure can be used to cluster time series based on the probability of co-occurrence of extreme events.
In this thesis, after a review of the math behind copulas, an application to the field of environmental sciences will be presented.
In the second chapter copulas and their main properties are mathematically defined, in particular Sklar's theorem and the invariance theorem are explained, these two theorems justifies the use of copulas to study the dependencies between the components of a random vector regardless of the marginals. Then three commonly used measures of association are introduced: the Spearman's rho and the Kendall's tau, which give an understanding of the general association, and the tail dependence coefficients which describe the dependence in the joint tails of bivariate distributions, so they are useful to study dependencies between extreme events. Finally after presenting the main copula families an overview of the main methods to estimate a copula and the statistical tests to select the most appropriate copula families is given.
In the third chapter methods to treat multivariate time series based on copulas are discussed. At first the concept of conditional copulas is introduced, then it is explained why it is possible to model each component of the time series individually, transform the residuals of the estimated models in pseudo-observations and apply on them the copula methods described in chapter two to study the dependencies between the components of the multivariate time series.
Finally a clustering method for time series, based on the measures of association defined in chapter two ,is presented. Starting from the estimated coefficients for each possible couple of components it is possible to define a dissimilarity function to perform a hierarchical clustering. Being based on copulas this clustering method is flexible and invariant under strictly increasing transformations of the time series.
In the Fourth chapter an application in the environmental field of all the methods previously discussed is shown. The time series of interest are made of the monthly rainfall maxima collected in various weather stations across the Friuli-Venezia Giulia region. The interest of this analysis lies in the co-occurrence of extreme events. After an exploratory analysis of the data, the marginal time series are modeled. Then the upper tail dependence coefficient for all the couples of stations are estimated, with these values a proper dis-

similarity matrix is computed and the stations for which the co-occurrence of extreme rainfall are most likely are clustered together. In the end it was possible to identify five different clusters of risk of co-occurrence of extreme events and it was possible to fit a suitable copula on each of them.

# Chapter 2

# Copulas

Copulas are used to study dependencies between random variables. Their usefulness follows directly from Sklar's theorem, the first part of the theorem states that for any d-dimensional distribution function $H$ with marginal univariate distributions $F_1, ..., F_d$ there exists a copula C such that:

$$H(\mathbf{x}) = C(F_1(x_1), ..., F_d(x_d)), \quad \mathbf{x} \in R^d. \tag{2.1}$$

This allows to study the marginal distributions separately from their joint dependence structure.
The mathematical explanation of copulas presented in this thesis will follow Joe [2], Nelsen [3] and Hofert, Kojadinovic, Mächler and Yan [8], using the same notation as in the latter.

## 2.1 Basic definitions and properties

### 2.1.1 Mathematical definition

The distribution function $H$ of a d-dimensional random variable $\mathbf{X} = (X_1, ...X_d)$ is defined as:

$$H(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x}) = P(X_1 \leq x_1, ..., X_d \leq x_d), \quad \mathbf{x} = (x_1, ..., x_d) \in R^d. \tag{2.2}$$

The distribution function $F_j$ of $X_j$, $j \in \{1, ..., d\}$ can be recovered from the multivariate distribution function H by $F_J(x_j) = H(\infty, ..., \infty, x_j, ..., \infty), x \in R$. $F_1, ...F_d$ are the marginal distribution functions of $\mathbf{X}$.
Supposing that all the marginals $F_i$ are continuous and applying the probability integral transform to each component of $\mathbf{X}$, a random vector

$\mathbf{U} = (F_1(X_1), ..., F_d(X_d)) = (U_1, ..., U_d)$ is obtained, this vector has uniform standard marginals, the copula of $\mathbf{X}$ is then defined as:

$$C(\mathbf{u}) = P(U_1 \leq u_1, ..., U_d \leq u_d), \quad \mathbf{u} = (u_1, ..., u_d) \in R^d. \tag{2.3}$$

So a copula is a d-dimensional distribution function on $[0,1]^d$ with standard uniform marginals and it represents the dependencies between the components of $\mathbf{X}$.

A function $C : [0,1]^d \to [0,1]$ is a copula if and only if:

1. C is grounded, meaning that $C(u_1, ..., u_d) = 0$ if $u_j$ is zero for at least one $j \in \{1, ..., d\}$.
2. C has standard uniform univariate margins, meaning that $C(1, ..., 1, u_j, 1..., 1) = u_j$ for all $u_j \in [0,1]$, $j \in \{1, ..., d\}$.
3. C is d-increasing, meaning that any C-volume $\Delta_{(\mathbf{a},\mathbf{b}]}C$ is nonnegative, for all $\mathbf{a} = (a_1, ..., a_d), \mathbf{b} = (b_1, ..., b_d) \in [0,1]^d, a_i \leq b_i$.

For any $\mathbf{a} = (a_1, ..., a_d), \mathbf{b} = (b_1, ..., b_d) \in [0,1]^d, \mathbf{a} \leq \mathbf{b}$, $(\mathbf{a}, \mathbf{b}]$ denotes the hyperrectangle defined by $\mathbf{u} \in [0,1]^d$:$\mathbf{a} < \mathbf{u} \leq \mathbf{b}$. For any hyperrectangle $(\mathbf{a}, \mathbf{b}]$ the C-volume is defined as:

$$\Delta(\mathbf{a}, \mathbf{b}]C = \sum_{\mathbf{i} \in \{0,1\}^d} (-1)^{\sum_{j=1}^d i_j} C(a_1^{i_1} b_1^{1-i_1}, \ldots, a_d^{i_d} b_d^{1-i_d}). \tag{2.4}$$

The simplest copula that can be defined is the independence copula:

$$\Pi(\mathbf{u}) = \prod_{j=1}^{d} u_j, \quad \mathbf{u} \in [0,1]^d, \tag{2.5}$$

which is the distribution function of a random vector $\mathbf{U} = (U_1, ..., U_d)$ with independent $U(0,1)$-distributed components; indeed in that case: $P(\mathbf{U} \leq \mathbf{u}) = P(U_1 \leq u_1, ..., U_d \leq u_d) = \prod_{j=1}^d P(\mathbf{U}_j \leq \mathbf{u}_j) = \prod_{j=1}^d u_j = \Pi(\mathbf{u})$.
In figure 2.1 it is possible to see on the left the surface plot and on the right the contour plot of the two-dimensional independence copula. From the surface plot it is possible to notice that $\Pi$ is zero on all edges of the unit square which start at $(0,0)$, and that $\Pi(u_1, 1) = u_1$ and $\Pi(1, u_2) = u_2$ for all $[u_1, u_2] \in [0,1]$, these are the necessary properties 1 and 2, as defined above, for a function to be a copula.
To show that also the property 3 is respected, in the bivariate case, it is first

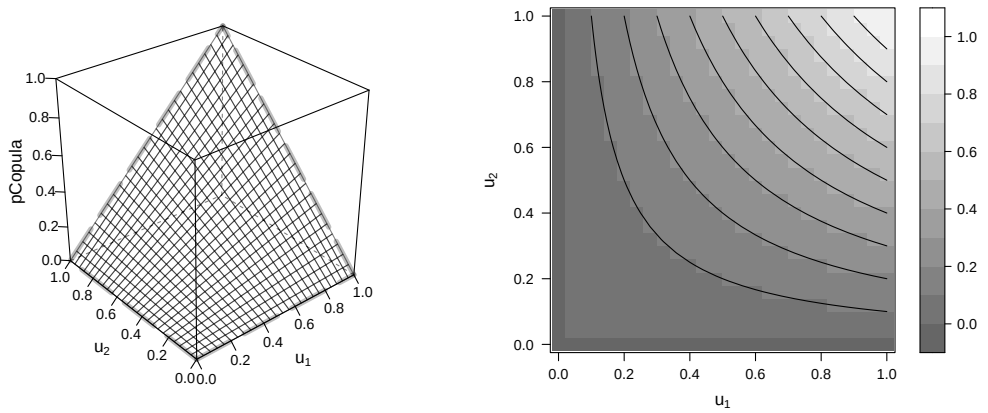


Figure 2.1: Surface plot (left) and contour plot(right) of the independence copula for d=2.

shown that the 2-d C-volume of $\Pi$ is:

$$\begin{aligned}\Delta(\mathbf{a},\mathbf{b}]\Pi &= \Pi(b_1,b_2) - \Pi(b_1,a_2) - \Pi(a_1,b_2) + \Pi(a_1,a_2)\\ &= b_1b_2 - b_1a_2 - a_1b_2 + a_1a_2\\ &= (b_1-a_1)(b_2-a_2).\end{aligned}$$

Then it is shown that:

$$\begin{aligned}P(\mathbf{U}\in(\mathbf{a},\mathbf{b}]) &= P(a_1 < U_1 \leq b1)P(a_2 < U_2 \leq b2)\\ &= (b_1-a_1)(b_2-a_2)\\ &= \Delta(a,b]\Pi.\end{aligned}$$

Since the C-volume is equal to a probability it follows that it is nonnegative. As it's shown in figure 2.2 the C-volume $\Delta(\mathbf{a},\mathbf{b}]\Pi$ can be approximated by the proportion of realizations of $U \sim \Pi$ falling in the hyperrectangle $(\mathbf{a},\mathbf{b}]$, this property holds true for all hyperrectangles and all copulas.
Finally a copula C is absolutely continuous if it admits a density, C admits a density c if:

$$c(\mathbf{u}) = \frac{\partial^d}{\partial u_d \cdots \partial u_1} C(u_1,\ldots,u_d), \quad \mathbf{u}\in(0,1)^d. \tag{2.6}$$

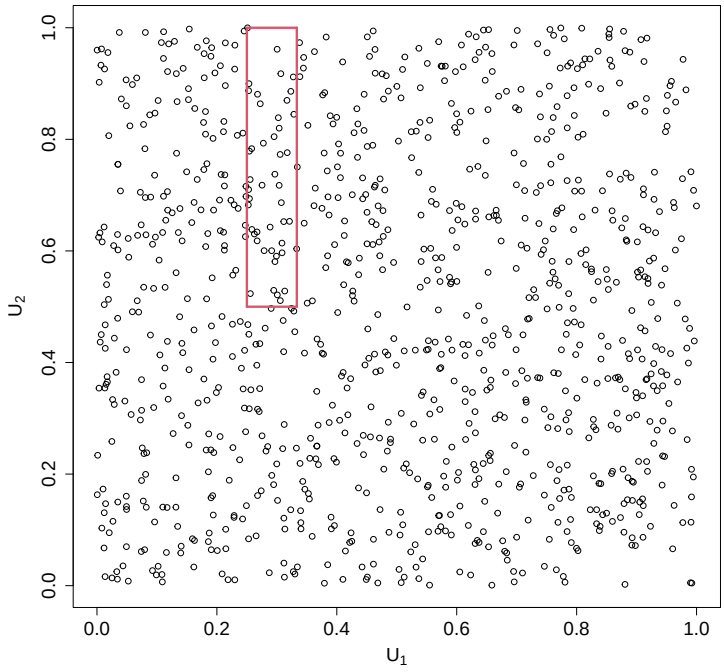


Figure 2.2: Scatter plot of n=1000 independent observations from Π and hyperrectangle with a=(1/4, 1/2) and b=(1/3, 1).

exists and is integrable.

### 2.1.2 The Fréchet–Hoeffding Bounds

Any d-dimensional copula C is pointwise bounded from below by the lower Fréchet–Hoeffding bound and from above by the upper Fréchet–Hoeffding bound:

$$W(\mathbf{u}) \leq C(\mathbf{u}) \leq M(\mathbf{u}), \quad \mathbf{u} \in [0,1]^d. \tag{2.7}$$

Where W and M are defined as:

$$W(\mathbf{u}) = \max\left\{\sum_{j=1}^{d} u_j - d + 1, 0\right\}, \quad M(\mathbf{u}) = \min_{1\leq j\leq d}\{u_j\}, \quad \mathbf{u} \in [0,1]^d. \tag{2.8}$$

W is a copula only if $d = 2$ and M is a copula only if $d \geq 2$.
If $U \sim U(0,1)$ then:

- W (only when $d = 2$) is the copula of the vector $(U, 1-U)$.
  W is called countermonotone copula and the dependence between the

components of $(U, 1-U)$ is referred to as perfect negative dependence, see the left plot of figure 2.3.

- M (for any $d \geq 2$) is the copula of the vector $(U, U, ..., U)$.
  M is called comonotone copula and the dependence between the components of $(U, U, ..., U)$ is referred to as perfect positive dependence, see the right plot of figure 2.3.

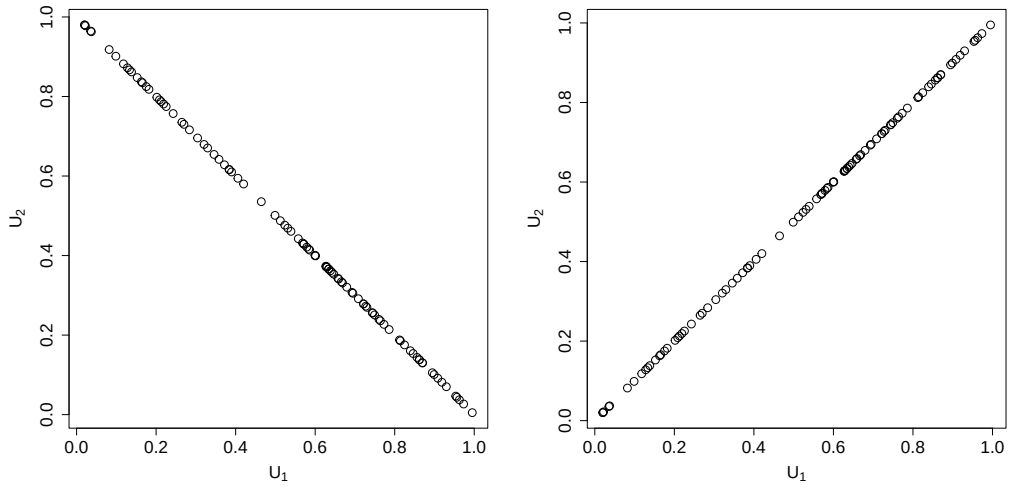


Figure 2.3: Scatter plot of n=1000 independent observations from W (left) and M (right) for d=2.

### 2.1.3 Sklar's theorem

Sklar's theorem is the most important result of copula theory, as it explains how copulas can describe the dependencies between the components of a random vector. In the following, given an univariate distribution function $F$, $ran\, F = \{F(x) : x \in R\}$ will denote the range of $F$, and $F^{\leftarrow}$ will denote the quantile function associated with F, which is the ordinary inverse $F^{-1}$ if $F$ is continuous and strictly increasing.

**Theorem 1 (Sklar's theorem)** *1. For any d-dimensional distribution function H with univariate margins $F_1, ..., F_d$, there exists a d-dimensional*

*copula $C$ such that:*

$$H(\boldsymbol{x}) = C(F_1(x_1), ..., F_d(x_d)), \quad \boldsymbol{x} \in R^d. \tag{2.9}$$

*The copula $C$ is uniquely defined on $ran\, F_1 \times ... \times ran\, F_d = \prod_j ran\, F_j$:*

$$C(\boldsymbol{u}) = H(F_1^{\leftarrow}(u_1), ..., F_d^{\leftarrow}(u_d)), \quad \boldsymbol{u} \in \prod_{j=1}^{d} ran\, F_j. \tag{2.10}$$

2. *Conversely, given a d-dimensional copula $C$ and univariate distribution functions $F_1, ..., F_d$, $H$ defined by equation 2.9 is a d-dimensional distribution function with margins $F_1, ..., F_d$.*

The first part of Sklar's theorem allows the decomposition of any d-dimensional distribution function $H$ into its univariate margins $F_1, ..., F_d$ and a copula C, essentially linking multivariate distribution functions to their univariate margins. Indeed if $\mathbf{X} = (X_1, ..., X_d) \sim H$ is a random vector with continuous margins $F_1, ..., F_d$, it holds that $U_i = F_i(X_i) \sim U(0, 1)$, if C denotes the distribution function of $U_1, ..., U_d$ then for any $\mathbf{x} \in \bar{R} = [-\infty, \infty]$:

$$\begin{aligned} H(x_1, ..., x_d) &= P(X_1 \leq x_1, ..., X_d \leq x_d) \\ &= P(F_1^{\leftarrow}(U_1) \leq x_1, ..., F_d^{\leftarrow}(U_d) \leq x_d) \\ &= P(U_1 \leq F_1(x_1), ..., U_d \leq F_d(x_d)) \\ &= C(F_1(x_1), ..., F_d(x_d)). \end{aligned}$$

If the margins are continuous then the copula is unique, otherwise it is uniquely defined on $ran\, F_1 \times ... \times ran\, F_d$. The fact that the underlying unknown copula is unique justifies its estimation from available data. If $\mathbf{X} \sim H$ with margins $F_j$ and equation 2.9 holds, it is said that $\mathbf{X}$ (or $H$) has copula $C$, the copula expresses the dependence on a quantile scale:

$$C(u_1, ..., u_d) = P(X_1 \leq F_1^{\leftarrow}(u_1), ..., X_1 \leq F_d^{\leftarrow}(u_d)).$$

The copula of $\mathbf{X}$ can be obtained by evaluating equation 2.9 at $x_i = F_i^{\leftarrow}(u_i)$, $0 \leq u_i \leq 1$, $i = 1, ..., d$:

$$\begin{aligned} C(u_1, ..., u_d) &= C(F_1(F_1^{\leftarrow}(u_1)), ..., F_d(F_d^{\leftarrow}(u_d))) \\ &= H(F_1^{\leftarrow}(u_1), ..., F_d^{\leftarrow}(u_d)). \end{aligned}$$

From the first part of the theorem it also follows that $H$ is absolutely continuous if and only if $C$ and $F_1, ..., F_d$ are absolutely continuous; in that case

the density $h$ of $H$ satisfies:

$$h(\mathbf{x}) = c(F_1(x_1), ..., F_d(x_d)) \prod_{j=1}^{d} f_j(x_j), \quad \mathbf{x} \in \prod_{j=1}^{d} ran\ X_j.$$

Where for any $j \in 1, ..., d$, $ran\, X_j = \{x \in R : P(X_j \in (x-h, x]) > 0$ for all $h > 0\}$ is the range of the random variable $X_j$, $f_j$ denotes the density of $F_j$ and $c$ denotes the density of $C$; $c$ can also be obtained from h via:

$$c(\mathbf{u}) = \frac{h(F_1^{\leftarrow}(u_1), ..., F_d^{\leftarrow}(u_d))}{f_1(F_1^{\leftarrow}(u_1)) \times ... \times f_d(F_d^{\leftarrow}(u_d))}, \quad \mathbf{u} \in (0,1)^d.$$

From the second part of the theorem it follows that new multivariate distribution functions can be constructed with given univariate margins and that copulas can be used to formulate dependence scenarios.
Finally two classes of distribution functions are defined.Considering $\mathbf{X} = (X_1, ..., X_d)$, a copula model

$$H(\mathbf{x}) = C(F_1(x_1), ..., F_d(x_d)), \quad \mathbf{x} \in R^d$$

can belong to:

1. The class of all multivariate distribution functions with given margins $F_1, ..., F_d$, which is the Fréchet class.
2. The class of all distribution functions obtained from a given d-dimensional copula $C$ known as meta-C models.

### 2.1.4 The invariance principle

**Theorem 2 (Invariance principle)** *Let $\boldsymbol{X} \sim H$ with continuous univariate margins $F_1, ..., F_d$ and a copula $C$. If, for any $j \in 1, ..., d$, $T_j$ is a strictly increasing transformation on $ran\, X_j$, then $T_1(X_1), ..., T_d(X_d)$ also has copula $C$.*

The invariance property allows the transformation of $\mathbf{X} = (X_1, ..., X_d)$ into $\mathbf{U} = (F_1(X_1), ..., F_d(X_d))$ without changing the underlying copula; the following lemma holds: $\mathbf{X}$ has copula $C$ if and only if $(F_1(X_1), ..., F_d(X_d)) \sim C$. Since $\mathbf{X}$ and $\mathbf{U}$ have the same copula it is possible to study the dependence between the components of $X$ by studying he dependence between the components of $U$ regardless of the marginals.
Two sampling algorithms follow directly from the lemma. The first one can be used to sample implicit copulas defined by equation 2.10:

1. Sample $\mathbf{X} \sim H$ where $H$ is a d-dimensional distribution function with continuous margins $F_1, ..., F_d$.
2. Return $\mathbf{U} = (F_1(X_1), ..., F_d(X_d))$.

The second one can instead be used to sample meta-C models:

1. Sample $\mathbf{U} \sim C$.
2. Return $\mathbf{X} = (F_1^{\leftarrow}(U_1), ..., F_d^{\leftarrow}(U_d))$.

### 2.1.5 Survival copulas and symmetries

Let $H(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x})$, $\mathbf{x} \in R^d$ be a multivariate distribution function, its corresponding multivariate survival function is defined as $\bar{H}(\mathbf{x}) = P(\mathbf{X} > \mathbf{x})$, $\mathbf{x} \in R^d$; generally the equivalence $\bar{H}(\mathbf{x}) = 1 - H(\mathbf{x})$ holds only when d=1 (univariate case).
Let $\mathbf{X}$ be a random vector with multivariate survival function $\bar{H}$, marginal distribution functions $F_i$ and hence marginal survival functions $\bar{F}_i = 1 - F_i,\ i \in \{1, ..., d\}$, it holds that:

$$\bar{H}(x_1, ..., x_d) = \bar{C}(\bar{F}_1(x_1), ..., \bar{F}_d(x_d)) \tag{2.11}$$

where $\bar{C}$ is the survival copula.
$\bar{C}$ is a copula and hence a distribution function, but $\bar{H}$ and $\bar{F}_1(x_1), ..., \bar{F}_d(x_d)$ are not distribution functions.
Finally let $C$ be a copula and let $\mathbf{U} \sim C$, then $1 - \mathbf{U} \sim \bar{C}$, i.e. $1 - \mathbf{U} = (1 - U_1, ..., 1 - U_d)$ is a random vector whose distribution function is the survival copula $\bar{C}$ corresponding to $C$. It follows that observations of $\mathbf{V} \sim \bar{C}$ can be obtained from observations of $\mathbf{U} \sim C$ using the relationship $\mathbf{V} = 1 - \mathbf{U}$. In figure 2.4 a bivariate Pareto–simplex copula, on the left, and the corresponding survival copula, on the right, can be seen, the point reflection with respect to the point $(1/2, 1/2)$ is noticeable.
There are two important kind of symmetries that can appear in multivariate distributions:

1. A random vector $\mathbf{X}$ is radially symmetric about $\mathbf{a} \in R^d$ if $\mathbf{X} - \mathbf{a} \overset{d}{=} \mathbf{a} - \mathbf{X}$, i.e. $\mathbf{X} - \mathbf{a}$ and $\mathbf{a} - \mathbf{X}$ are equal in distribution.
   Moreover if $X_j$ is symmetric about $a_j$ then $\mathbf{X}$ is radially symmetric about $\mathbf{a}$ iff $C = \bar{C}$, $C$ is radially symmetric, see left figure 2.5.
2. A random vector $\mathbf{X}$ is exchangeable if $(X_{j1}, ..., X_{jd}) \overset{d}{=} (X_1, ..., X_d)$ for all permutations $(j_1, ..., j_d)$ of $\{1, ..., d\}$.
   Moreover if $C(u_{j1}, ..., u_{jd}) = C(u_1, ..., u_d)$ for all $u_1, ..., u_d \in [0, 1]$ and

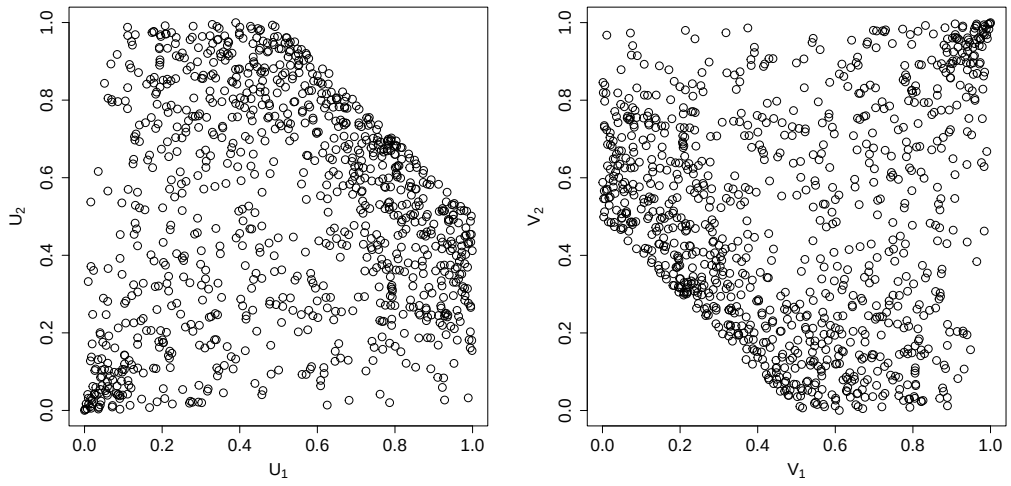


Figure 2.4: Scatter plot of n=1000 independent observations from a Pareto–simplex copula (left) and survival Pareto–simplex copula (right).

all permutations $(j_1, ..., j_d)$ of $\{1, ..., d\}$, $C$ is exchangeable, see right figure 2.5.

## 2.2 Measures of association

In the bivariate case one of the most used measure of association is the Pearson's (or linear) correlation coefficient, defined for a random vector $(X_1, X_2)$, whose components have finite variances, by:

$$\begin{aligned} Cor(X_1, X_2) &= \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1)}\sqrt{Var(X_2)}} = \\ &= \frac{E((X_1 - E(X_1))(X_2 - E(X_2)))}{\sqrt{E((X_1 - E(X_1))^2)}\sqrt{E((X_2 - E(X_2))^2)}} \end{aligned} \tag{2.12}$$

But the Pearson's coefficient is merely a measure of linear dependence and its usefulness is truly meaningful only in the context of the so-called elliptical distributions such as the multivariate normal or t distributions [8].
In this section other measures of association will be defined, in particular

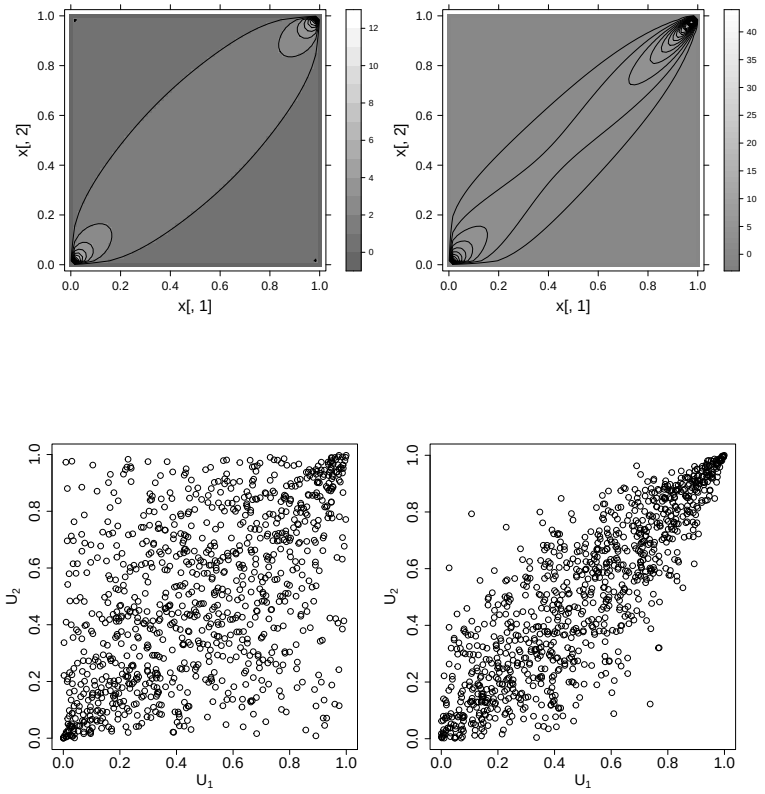


Figure 2.5: Contour plot and observations of bivariate t-copula ($\rho = 0.5$ $\nu = 2.5$), both exchangeable and radially symmetric (left), and Contour plot and observations of Gumbel-Hougaard ($\theta = 3$), only exchangeable (right)

these measures will only depend on the underlying unique copula and so they will be more useful than the linear coefficient.

### 2.2.1 Rank correlation measures

Let $(X_1, X_2)$ be a bivariate random vector with continuous marginal distribution functions $F_1$ and $F_2$, the two most used rank correlation measures are:

1. The (population version of) Spearman's rho, which is defined as:

$$\rho_s = \rho_s(X_1, X_2) = cor(F_1(X_1), F_2(X_2)). \tag{2.13}$$

2. Let $(X_1', X_2')$ be an independent copy of $(X_1, X_2)$, the (population ver-

sion of) Kendall's tau is defined as:

$$\tau = \tau(X_1, X_2) = E\big[sign((X_1 - X_1')(X_2 - X_2'))\big]. \tag{2.14}$$

Both the measures can be seen as measures of concordance; given $(X_1, X_2)$ and an independent copy $(X_1', X_2')$, the vectors are concordant if $(X_1 - X_1')(X_2 - X_2') > 0$ and discordant if the opposite holds. The Kendall's tau can be rewritten as:

$$\tau = P((X_1 - X_1')(X_2 - X_2') > 0) - P((X_1 - X_1')(X_2 - X_2') < 0) \tag{2.15}$$

which is the probability of concordance minus the probability of discordance. Analogously the Spearman's rho can be rewritten as:

$$\rho_s = 3[P((X_1 - X_1')(X_2 - X_2') > 0) - P((X_1 - X_1')(X_2 - X_2') < 0)] \tag{2.16}$$

Finally Both of the measures can be expressed exclusively in terms of the underlying copula, indeed if $(X_1, X_2)$ is a bivariate random vector with continuous marginal distribution functions and copula C, the Spearman's rho can be written as:

$$\rho_s = \rho_s(C) = 12 \int_{[0,1]^2} C(\mathbf{u}) d\mathbf{u} - 3 = 12 \int_{[0,1]^2} u_1 u_2 dC(\mathbf{u}) - 3 \tag{2.17}$$

and the Kendall's tau as:

$$\tau = \tau(C) = 4 \int_{[0,1]^2} C(\mathbf{u}) dC(\mathbf{u}) - 1. \tag{2.18}$$

From these definitions it follows that Spearman's rho and Kendall's tau can be considered as the moments of the copula.
Given a random sample $(X_{11}, X_{12}), ..., (X_{n1}, X_{n2})$ it is possible to estimate the two measures. Given the sample of bivariate ranks $(R_{11}, R_{12}), ..., (R_{n1}, R_{n2})$, where $R_{ij}$ is the rank of $X_{ij}$ among $X_{1j}, ..., X_{nj}$, the Spearman's rho is estimated as:

$$\rho_{s,n} = \frac{\sum_{i=1}^{n}(R_{i1} - \bar{R}_1)(R_{i2} - \bar{R}_2)}{\sqrt{\sum_{i=1}^{n}(R_{i1} - \bar{R}_1)^2}\sqrt{\sum_{i=1}^{n}(R_{i2} - \bar{R}_2)}}, \tag{2.19}$$

where $\bar{R}_1 = \bar{R}_2 = (n+1)/2$ is the mean rank of the two component series. The sample version of the Kendall's tau can be estimated as:

$$\tau_n = \frac{4p_n}{n(n-1)} - 1 \tag{2.20}$$

with $p_n$ number of concordant pairs in the sample.

As briefly explained in Genest Favre [9], most of other used correlation measures are based upon expressions of the form:

$$\int J(u_1, u_2) dC(u_1, u_2)$$

where $J$ is a proper score function, for example in the Spearman's rho case $J(u_1, u_2) = u_1 u_2$.

### 2.2.2 Tail dependence coefficients

Rank correlation measures are not very suitable to study the dependencies between extreme events, so tail dependence coefficients are introduced, they are used to describe the dependence in the joint tails of bivariate distributions, and they are defined as limits of the conditional probabilities of quantile exceedances.
Let $(X_1, X_2)$ be a random vector with continuous marginal distribution functions $F_1$ and $F_2$ and copula $C$:

1. The lower tail dependence coefficient of $X_1$ and $X_2$ is defined as:

$$\lambda_l = \lambda_l(X_1, X_2) = lim_{q \to 0^+} P(X_2 \leq F_2^{\leftarrow}(q) | X_1 \leq F_1^{\leftarrow}(q)). \tag{2.21}$$

2. The upper tail dependence coefficient of $X_1$ and $X_2$ is defined as:

$$\lambda_u = \lambda_u(X_1, X_2) = lim_{q \to 1^-} P(X_2 > F_2^{\leftarrow}(q) | X_1 > F_1^{\leftarrow}(q)). \tag{2.22}$$

As in the rank correlation case, the tail dependence coefficients can be expressed exclusively in terms of the underlying copula $C$, the lower tail dependency coefficient can be written as:

$$\lambda_l = \lambda_l(C) = lim_{q \to 0^+} \frac{P(X_2 \leq F_2^{\leftarrow}(q), X_1 \leq F_1^{\leftarrow}(q))}{P(X_1 \leq F_1^{\leftarrow}(q))} = lim_{q \to 0^+} \frac{C(q, q)}{q}. \tag{2.23}$$

and the upper tail dependence coefficient can be written as

$$\begin{aligned} \lambda_u = \lambda_u(C) = lim_{q \to 0^+} \frac{\bar{C}(q, q)}{q} = lim_{q \to 1^-} \frac{\bar{C}(1-q, 1-q)}{1-q} = \\ = lim_{q \to 1^-} \frac{1 - 2q + C(q, q)}{1-q}. \end{aligned} \tag{2.24}$$

Where $\bar{C}$ is the previously defined survival copula. From these definitions it follows that radially symmetric copulas have the same upper and lower tail dependence coefficients, since in that case $C = \bar{C}$.
The estimation of the tail dependence coefficient is more complex than that of the rank measures; as presented in Frahm, Junker, Schmidt [4] there are several methods of estimation for the tail dependence coefficient, a commonly used nonparametric estimator is based on the empirical copula, which will be defined later in equation 2.38, and is defined as:

$$\hat{\lambda}_U^{LOG} = 2 - \frac{C_n((n-k)/n, (n-k)/n)}{log((n-k)/n)} \quad 0 < k < n, \tag{2.25}$$

where $C_n(\mathbf{u})$ is the empirical copula, n the number of observations and k is the percentile threshold considered for the tail.

## 2.3 Copula examples

In this section the most common copula families will be introduced, some, like the elliptical copulas, are implicit, and are obtained by applying Sklar's theorem to common multivariate distributions (such as the elliptical ones), others, like the archimedean copulas, are explicit, and have simple closed form.
The last family presented is that of the extreme-value copulas, which are the copulas of multivariate extreme-value distributions.

### 2.3.1 Elliptical copulas

These are the copulas of elliptical distributions, such as the gaussian and the t-distribution.
Let $\mathbf{Y} \sim N_d(\mu, \Sigma)$, then its copula is the same as the copula of $\mathbf{X} \sim N_d(\mu, P)$, where $P$ is the correlation matrix of $\mathbf{Y}$. The gaussian copula family is then defined as:

$$\begin{aligned} C_P^{Ga}(\mathbf{u}) &= P(\Phi(X_1) \leq u_1, ..., \Phi(X_d) \leq u_d) \\ &= P(X_1 \leq \Phi^{-1}(u_1), ..., X_d \leq \Phi^{-1}(u_d)) \\ &= \mathbf{\Phi}_P(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_d)). \end{aligned}$$

Where $\mathbf{\Phi}_P$ is the joint distribution of $\mathbf{X}$ and $\Phi$ is the cumulative distribution of $N(0,1)$, from the choice of the correlation matrix follows some properties:

- if $d = 2$ then $P = \rho = corr(X_1, X_2)$.

- $P = I_d$ gives the independence copula.
- if P is a matrix of ones, then C is the comonotonicity copula.
- if $d = 2$ and $\rho = -1$ then C is the countermonotonicity copula.

In figure 2.6 it is possible to see some examples of $2-d$ gaussian copulas, with respectively $\rho = 0.1$, $\rho = 0.5$, $\rho = 0.95$, it can be noticed that for $\rho = 0.1$ the gaussian copula tends to the independent copula, while for $\rho = 0.95$ it tends to the comonotonicity copula, moreover gaussian copulas are both exchangeable and radially symmetric. The t copula family $C^t_{P,\nu}$ is obtained applying

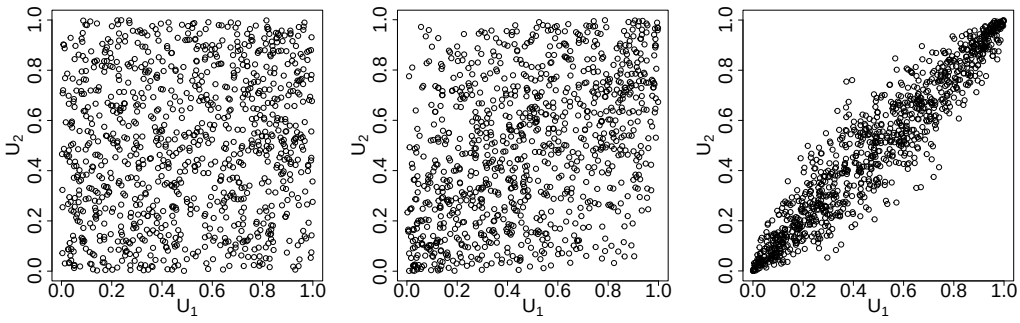


Figure 2.6: Scatter plots of n=1000 independent observations from $C^{Ga}_{\rho}$ with $\rho = 0.1$, $\rho = 0.5$, $\rho = 0.95$.

the Sklar's theorem to a multivariate t distribution $t_{P,\nu}$, with location vector 0, scale matrix P, and $\nu > 0$ degrees of freedom:

$$C^t_{P,\nu}(u) = t_{P,\nu}(t_\nu^{-1}(u_1), ..., t_\nu^{-1}(u_d))$$
$$= \int_{-\infty}^{t_\nu^{-1}(u_d)} \cdots \int_{-\infty}^{t_\nu^{-1}(u_1)} \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})(\pi\nu)^{d/2}\sqrt{detP}} \Big(1 + \frac{\mathbf{x}' P^{-1}\mathbf{x}}{\nu}\Big)^{-\frac{\nu+d}{2}} dx_1 ... dx_d,$$

where $t_\nu^{-1}$ is the quantile function of the distribution $t_\nu$ of the univariate Student t distribution with $\nu$ degrees of freedom, from the choice of the scale matrix follows some properties:

- if $d = 2$ then $C^t_{-1,\nu}$ is the countermonotonicity copula.
- if $d \geq 2$ and P is a matrix of ones, $C^t_{P,\nu}$ is the comonotonicity copula.
- P=$I_d$ does not give the independence copula.

In figure 2.7 it is possible to see some examples of $2-d$ t copulas, above with fixed $\nu = 2.5$ and with respectively $P = -0.99$, $P = 0.5$, $P = 0.99$, it can be noticed that for $P = -0.99$ the copula tends to the countermonotonicity copula and for $P = 0.99$ the copula tends to the comonotonicity copula; below with fixed $P = 0.5$ and with respectively $\nu = 1$, $\nu = 3$, $\nu = 5$, it can be seen that a lower degree of freedom results in a bigger tail dependence coefficient. Finally all t copulas are both exchangeable and radially symmetric.

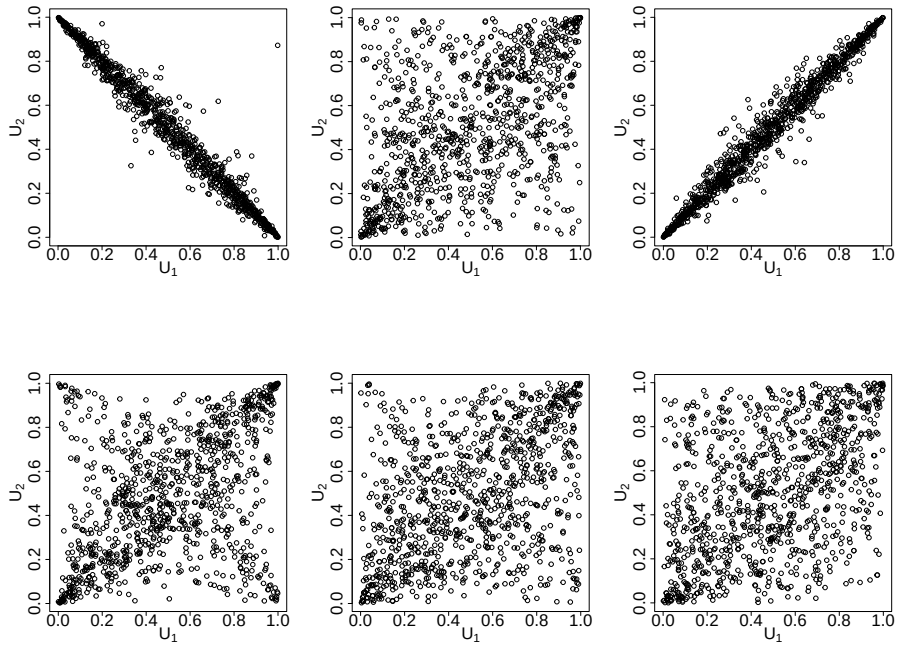


Figure 2.7: Above scatter plots of n=1000 observations from $C^t_{P,\nu}$ with fixed $\nu = 2.5$ and $P = -0.99$, $P = 0.5$, $P = 0.99$. Below scatter plots of n=1000 observations from $C^t_{P,\nu}$ with fixed $P = 0.5$ and $\nu = 1$, $\nu = 3$, $\nu = 5$.

### 2.3.2 Archimedean copulas

A copula is archimedean if it can be written in the form:

$$C(\mathbf{u}) = \psi(\psi^{-1}(u_1) + ... + \psi^{-1}(u_d)), \quad \mathbf{u} \in [0,1]^d. \tag{2.26}$$

where $\psi$ is a function known the generator of $C$, and $\psi^{-1}$ is its pseudoinverse. It follows from their definition that archimedean copulas are exchangeable. Some of the most often used archimedean copulas, defined through a parametric generator $\psi_\theta(t)$, are:

- The Gumbel-Hougaard copula ($d = 2$)

$$C_\theta^{Gu}(u_1, u_2) = exp(-((-log(u_1))^\theta + (-log(u_2))^\theta)^{1/\theta}), \quad \theta \geq 1.$$

  $\theta = 1$ gives the independence copula and $\theta \to \infty$ gives the comononotonicity copula.

- The Clayton copula ($d = 2$)

$$C_\theta^{C}(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}, \quad \theta > 0.$$

  $\theta \to 0$ gives the independence copula and $\theta \to \infty$ gives the comononotonicity copula.

- The Frank copula

$$C_\theta^{F}(u_1, u_2) = -\frac{1}{\theta} log\Big(1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1}\Big).$$

  $\theta \to 0$ gives the independence copula and $\theta \to \infty$ gives the comononotonicity copula.

These kind of copulas are particularly useful because they allow to model the dependence through the single parameter $\theta$.
In figure 2.8 it is possible to see the plots of the Gumbel-Hougaard, Clayton and Frank copula with respectively parameter $\theta$ equal to 9, 3 and 3; the different tail dependencies are noticeable.

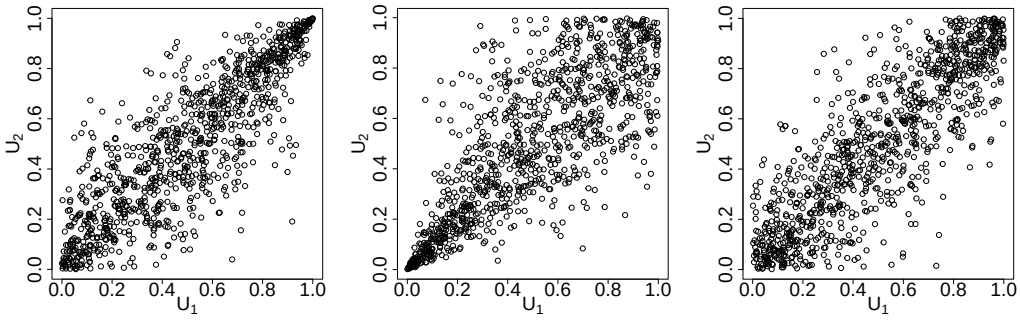


Figure 2.8: Scatter plots of n=1000 independent observations from Gumbel-Hougaard, Clayton and Frank copula with respectively parameter $\theta$ equal to 9, 3 and 3.

### 2.3.3 Extreme value copulas

These are the copulas of multivariate extreme-value distributions. A copula is an extreme-value copula if there exists a copula $C^*$ such that for any $\mathbf{u} \in [0,1]^d$:

$$lim_{n\to\infty} C^*(\mathbf{u}^{1/n})^n = C(\mathbf{u}), \tag{2.27}$$

$C$ belongs to the domain of attraction of $C^*$.
Moreover it also holds that a copula is an extreme-value one if and only if it is max-stable, which means that for any $\mathbf{u} \in [0,1]^d$ and $r \in N$ it holds that $C(\mathbf{u}^{1/r})^r = C(\mathbf{u})$.
The previously defined Gumbel-Hougaard copula belongs also to the extreme-value family, other commonly used extreme-value copulas are the Galambos, the Tawn and the t-EV.

## 2.4 Estimation and goodness of fit

Let $\mathbf{X}_1, ..., \mathbf{X}_n$ be a random independent sample of a d-dimensional random vector $\mathbf{X}$ with distribution function H and marginal distributions $F_1, ..., F_d$, it follows from Sklar's theorem that there exists a unique copula $C$ on $[0,1]^d$ such that:

$$H(\mathbf{x}) = C(F_1(X_1), ..., F_d(X_d)) \quad \mathbf{x} \in R^d$$

In this section some methods to estimate the copula $C$ starting from the samples, and some of the statistical tests to choose the most appropriate copula family, will be presented.

### 2.4.1 Estimation

It is assumed that the copula $C$ belongs to an absolutely continuous parametric family of copulas:

$$\mathbf{C} = \{C_\theta : \theta \in \Theta\}$$

where $\Theta$ is the parameter space, a subset of $R^p$ with integer $p \geq 1$; if $C \in \mathbf{C}$ then there must exists a $\theta_0 \in \Theta$ such that $C = C_{\theta_0}$, so estimating the copula is equivalent to estimating the parameter vector $\theta_0$.
At first it is assumed that the univariate margins $F_1, ..., F_d$ belong to an absolutely continuous parametric families of distribution functions:

$$\mathbf{F_j} = \{F_{j,\gamma_j} : \gamma_j \in \Gamma_j\},$$

where $\Gamma_j$ is a subset of $R^{p_j}$ for an integer $p_j > 0$, so for any $j \in \{1, ..., d\}$ there exists a $\gamma_{0,j} \in \Gamma_j$ such that $F_j = F_{j,\gamma_{0,j}}$.
The problem so is estimating the copula $C_{\theta_0}$ from a sample $\mathbf{X_1}, ...\mathbf{X_n}$ assuming that the distribution function $H$ belongs to the parametric family:

$$\mathbf{H} = \Big\{ C\big(F_1(.), ..., F_d(.)\big) : C \in \mathbf{C} \, and \, F_j \in \mathbf{F_j} \, for \, all \, j \in \{1, ..., d\} \Big\}.$$

Since both $\mathbf{C}$ and all $\mathbf{F_J}$ are absolutely continuous, also $\mathbf{H}$ is absolutely continuous.
Given $H \in \mathbf{H}$ the corresponding density h is:

$$h(\mathbf{x}) = c_\theta\big(F_{1,\gamma_1}(x_1), ..., F_{d,\gamma_d}(x_d)\big) \prod_{j=1}^{d} f_{j,\gamma_j}(x_j),$$

Where $c_\theta$ is the density of $C_\theta$ and $f_{j,\gamma_j}$ is the density of $F_{j,\gamma_j}$, $j \in \{1, ..., d\}$. The parameter vector $(\gamma_{0,1}, ..., \gamma_{0,d}, \theta_0)$ can be estimated by the Maximum Likelihood Estimator:

$$l_n(\gamma_1, ..., \gamma_d, \theta) = \sum_{i=1}^{n} log \, c_\theta\big(F_{1,\gamma_1}(X_{i1}), ..., F_{d,\gamma_d}(X_{id})\big) + \sum_{j=1}^{d} \sum_{i=1}^{n} log \, f_{j,\gamma_j}(X_{ij}). \tag{2.28}$$

This gives the estimate $C_{\theta_n}$ of $C$ and also the estimate $F_{j,\gamma_{n,j}}$ of $F_{j,\gamma_j}$, consequently giving also an estimate $C_{\theta_n}\big(F_{1,\gamma_{n,1}}(.), ..., F_{d,\gamma_{n,d}}(.)\big)$ of $H$.

To reduce the complexity of the maximum likelihood estimation it is possible to implement a two stage estimator, this method is known as inference functions for margin estimator. At first each univariate margins is estimated via:

$$\gamma_{n,j} = argsup_{\gamma_{n,j} \in \Gamma_j} \sum_{i=1}^{n} log \, f_{j,\gamma_j}(X_{ij}), \tag{2.29}$$

the estimated margins are then used to compute a sample of the so called parametric pseudo-observations from C:

$$\mathbf{U}_{i,\gamma_n} = \big(F_{1,\gamma_{n,1}}(X_{i1}), ..., F_{d,\gamma_{n,d}}(X_{id})\big), \quad i \in \{1, ..., n\}. \tag{2.30}$$

The parametric pseudo-observations are then used to estimate $\theta_0$ maximizing a log-likelihood function:

$$\theta_n = argsup_{\theta \in \Theta} \sum_{i=1}^{n} log \, c_\theta(\mathbf{U}_{i,\gamma_n}). \tag{2.31}$$

These two methods can lead to biased estimation of $\theta_0$ if the margins are misspecified, to avoid this problem it is possible to estimate them nonparamatrically through the rescaled empirical distribution functions of the component samples of $\mathbf{X_1}, ..., \mathbf{X_n}$:

$$F_{n,j} = \frac{1}{n+1} \sum_{i=1}^{n} 1(X_{ij} \leq x), \quad x \in R, \tag{2.32}$$

which are used to compute the (nonparametric) pseudo-observations sample:

$$\mathbf{U}_{i,n} = \big(F_{n,1}(X_{i1}), ..., F_{n,d}(X_{id})\big), \quad i \in \{1, ..., n\}, \tag{2.33}$$

which can then be used to estimate $\theta_\mathbf{0}$ via the maximum pseudo-likelihood estimator:

$$\theta_n = argsup_{\theta \in \Theta} \sum_{i=1}^{n} log\, c_\theta(\mathbf{U}_{i,n}). \tag{2.34}$$

Finally if $R_{ij}$ is the rank of $X_{ij}$ among $X_{1j}, ..., X_{nj}$ then $F_{n,j}(X_{ij}) = R_{ij}/(n+1)$, i.e. the pseudo-observations are a sample of multivariate scaled ranks:

$$\mathbf{U}_{i,n} = \frac{1}{n+1}(R_{i1}, ..., R_{id}), \quad i \in \{1, ..., n\}. \tag{2.35}$$

Another possibility is to use the method of moments estimator, in particular using the previously defined moments of the copula, the Kendall's tau and the Spearman's rho. In the bivariate case given a copula C, the function $g_\tau$ and $g_{\rho_s}$ are defined as:

$$g_\tau(\theta) = \tau(C_\theta) \qquad g_{\rho_s} = \rho_s(C_\theta), \quad \theta \in \Theta \subseteq R, \tag{2.36}$$

if the functions $g$ in the equations 2.36 are injective then it is possible to use the method of estimators, the estimator $\theta_n$ of $\theta_0$ is:

$$\theta_n = g_\tau^{-1}(\tau_n) \quad or \quad \theta_n = g_{\rho_s}^{-1}(\rho_{s,n}), \tag{2.37}$$

All the variables used here are defined in the section 2.2.1.
If the dimension $d$ is greater than two, the copula C is exchangeable and there is only one parameter, it is possible to use the method of moments estimator by applying the function g to the average of the sample tau (or rho) of $\binom{d}{2}$ different bivariate margins.

Finally a nonparametric estimator of the copula is the empirical copula of $\mathbf{X_1}, ..., \mathbf{X_n}$:

$$C_n(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} 1(\mathbf{U_{i,n}} \leq \mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d} 1(U_{i,j,n} \leq u_j), \quad \mathbf{u} \in [0,1]^d. \quad (2.38)$$

This is the empirical distribution function of the pseudo-observations. It is a consistent estimator of $C$, and its asymptotics follow from the empirical copula process:

$$\sqrt{n}(C_n(\mathbf{u}) - C(\mathbf{u})), \quad \mathbf{u} \in [0,1]^d \quad (2.39)$$

### 2.4.2 Goodness of fit

Some graphical and statistical procedures can be used to choose the best parametric copula family.
In the bivariate case a graphical test can be done by simply plotting the scatterplot of the pseudo-observations, if the dimension d is greater than two, but not too big, the test can be done on all the $\binom{d}{2}$ different bivariate margins. With this approximate method it is possible to recognize wether the copula belongs to one of the common families previously defined, by looking at some properties such as symmetries and tail dependencies.

After this analysis more rigorous tests are required to confirm the hypothesis made about the properties of the copula, considering the sample $\mathbf{X}_1, ..., \mathbf{X}_n$, there are test statistics based on the empirical copula, some of the most helpful tests are:

- test of independence, to check wether the copula is independent, with test statistic:
$$S_n^{\Pi} = \int_{[0,1]^d} n(C_n(\mathbf{u}) - \Pi(\mathbf{u}))^2 d\mathbf{u}. \quad (2.40)$$

- Test of exchangeability, to check wether the copula is exchangeable, with test statistic, in the bivariate case:
$$S_n^{exc} = \int_{[0,1]^2} n(C_n(u_1, u_2) - C_n(u_2, u_1))^2 dC_n(\mathbf{u}) \quad (2.41)$$

- Test of radial symmetry, to check wether the copula is radially symmetric, with test statistic:
$$S_n^{sym} = \int_{[0,1]^d} n(C_n(\mathbf{u}) - \bar{C}_n(\mathbf{u}))^2 dC_n(\mathbf{u}) \quad (2.42)$$

- Test of extreme-value dependence, to check wether the copula belongs to the extreme-value family, with test statistic:

$$S_n^{evC} = T_{3,n} + T_{4,n} + T_{5,n}, \tag{2.43}$$

where:

$$T_{r,n} = \int_{[0,1]^d} n\left(\left(C_n(u_1^{1/r}, ..., u_d^{1/r})\right)^r - C_n(\mathbf{u})\right)^2 dC_n(\mathbf{u}).$$

Once the choice of the parametric family is limited to few options, based on results of the aforementioned tests, for each of the remaining family $\mathbf{C}$ a goodness of fit test can be run. The goodness of fit tests the hypothesis:

$$H_0 : C \in \mathbf{C} \quad versus \quad H_1 : C \notin \mathbf{C}.$$

One of the most common goodness of fit tests is comparing the empirical copula $C_n$ with an estimate $C_{\theta_n}$ of $C$ obtained under the assumption that $C \in \mathbf{C}$. $\theta_n$ is an estimator of $\theta$, computed from the pseudo-observations via the maximum pseudo-likelihood estimator. The Cramér-von Mises statistic is defined as:

$$S_n^{gof} = \int_{[0,1]^d} n\big(C_n((u) - C_{\theta_n}(\mathbf{u}))\big)^2 dC_n(\mathbf{u}) = \sum_{i=1}^{n} \big(C_n(\mathbf{U}_{i,n} - C_{\theta_n}(\mathbf{U}_{i,n}))\big)^2. \tag{2.44}$$

Using a parametric bootstrap it is then possible to obtain an approximate p-value for this test, for the algorithm see Hofert, Kojadinovic, Mächler and Yan [8].

# Chapter 3

# Copula methods for time series

Consider a random d-dimensional vector $\mathbf{X}$ that is observed at successive points in time, i.e. the observations $(\mathbf{X}_i)_{i \in Z}$ form a time series.
In the time series settings what will be modeled is the conditional distribution of $\mathbf{X}_i$ given the information of past values $\mathbf{X}_1, ..., \mathbf{X}_{i-1}$.
Copula methods to treat multivariate time series, such as the concept of the conditional copula obtained by applying Sklar's theorem to the conditional distribution of $\mathbf{X}_i$ given its past values, were first introduced by Patton [5], and further expanded in his successive works (see [10] and [6]), and in the work of Rémillard [11].
Finally a copula based method for clustering time series will be introduced.
As in the previous chapter the structure and the notation used will follow Hofert, Kojadinovic, Mächler and Yan[8].

## 3.1 Conditional copulas

In a general case, if $(\mathbf{X}, \mathbf{Z})$ is a $(d+q)$-dimensional vector, and $ran\,\mathbf{Z} = \{\boldsymbol{z} \in R^q : P\big(\boldsymbol{z} \in (\boldsymbol{z} - \mathbf{h}, \boldsymbol{z}]\big) > 0 \; for\, all\, \mathbf{h} > 0\big)\}$ is the range of the random vector $\mathbf{Z}$, then the conditional distribution of $\mathbf{X}$ given $\mathbf{Z} = \boldsymbol{z}$ is defined as:

$$H_z(\mathbf{x}) = P(X_1 \leq x_1, ..., X_d \leq x_d | \mathbf{Z} = \boldsymbol{z}), \quad \mathbf{x} \in R^d. \tag{3.1}$$

If $H_z$ is continuous it follows from Sklar's theorem that:

$$H_z(\mathbf{x}) = C_z(F_{\boldsymbol{z},1}(x_1), ..., F_{\boldsymbol{z},1}(x_d)), \quad \mathbf{x} \in R^d, \tag{3.2}$$

and that:

$$C_z(\mathbf{u}) = P(F_{\boldsymbol{z},1}(X_1) \leq u_1, ..., F_{\boldsymbol{z},d}(X_d) \leq u_d | \mathbf{Z} = \boldsymbol{z}), \quad \mathbf{u} \in [0,1]^d. \tag{3.3}$$

Where $F_{\boldsymbol{z},j}(x) = P(X_j \leq x | \mathbf{Z} = \boldsymbol{z})$, $j \in \{1, ..., d\}$ are the univariate margins of $H_z$ and $C_z$ is its copula.

In the time series settings the information set $G_{i-j}$ generated by previous observations $\{\mathbf{X}_{i-j},\ j = 1, 2, ...\}$ is considered, so the aim is to model:

$$H_{G_{i-1}}(\mathbf{x}) = P(X_{i1} \leq x_1, ..., X_{id} \leq x_d | G_{i-1}), \quad \mathbf{x} \in R^d,$$

which is the conditional distribution of $\mathbf{X}_i$ given the information $G_{i-1}$ at time $i-1$.
If $F_{G_{i-1},j}(x) = P(X_{ij} \leq x | G_{i-1})$, $x \in R$, $j \in \{1, ..., d\}$ are the margins, under continuity and measurability assumption it holds that:

$$H_{G_{i-1}}(\mathbf{x}) = C_{G_{i-1}}(F_{G_{i-1},1}(x_1), ..., F_{G_{i-1},1}(x_d)), \quad \mathbf{x} \in R^d, \tag{3.4}$$

where $C_{G_{i-1}}$ is the conditional copula.
The same information set must be used in each of the marginals and for the copula in order for the resulting function to be a multivariate conditional joint distribution. However some of the information contained in $G_{i-1}$ could be not relevant for all variables, for example, it might be that each variable depends on its own first lag, but not on the lags of any other variable; consider $G_{i-1,j}$ as the smallest subset of $G_{i-1}$ such that $X_{ij}|G_{i-1,j}$ has the same distribution as $X_{ij}|G_{i-1}$; with this it is possible to construct each marginal distribution model using only $G_{i-1,j}$, and then use $G_{i-1}$ for the copula, to obtain a valid conditional joint distribution [10].

## 3.2 Marginals modeling

It is assumed that all the marginal distributions have the form:

$$X_{ij} = \mu_{ij}(\boldsymbol{\beta}_j) + \sigma_{ij}(\boldsymbol{\beta}_j)\epsilon_{ij}, \tag{3.5}$$

where $\mu_{ij}(\boldsymbol{\beta}_j) = E(X_{ij}|G_{i-1})$ and $\sigma_{ij}(\boldsymbol{\beta}_j) = Var(X_{ij}|G_{i-1})$ are respectively the conditional mean and variance of $X_{ij}$ given $G_{i-1}$; furthermore, for any $j \in \{1, ..., d\}$, the conditional distribution of the innovations $\epsilon_{ij}$ given $G_{i-1}$ does not depend on $G_{i-1}$ and it has mean equal to zero, variance equal to one and its distribution functions belong to an absolutely parametric family $\mathbf{F_j} = \{F_{j,\gamma_j} : \gamma_j \in \Gamma_j\}$.

On the contrary the conditional distribution of $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{id})$ given $G_{i-1}$ can depend on $G_{i-1}$, meaning that while the univariate margins of $\boldsymbol{\epsilon}_i$ given $G_{i-1}$ do not depend on $G_{i-1}$, the conditional copula of $\boldsymbol{\epsilon}_i$ given $G_{i-1}$ can depend on $G_{i-1}$. It is also assumed that there exists an absolutely continuous parametric family of copula $\mathbf{C} = \{C_\theta : \theta \in \Theta\}$, and a parametric copula calibration function $\boldsymbol{\phi}$ such that the conditional copula of $\boldsymbol{\epsilon}_i$ given $G_{i-1}$ is $C_{\boldsymbol{\phi}(G_{i-1})} = C_{\theta_{i-1}}$. Finally the conditional mean and variance are defined up to a finite dimensional parameter vector $\boldsymbol{\beta}_j$, and $\phi$ is defined up to a finite dimensional parameter vector $\boldsymbol{\beta}$.
Putting all these together it follows that the conditional distribution function of $\boldsymbol{\epsilon}_i$ given $G_{i-1}$ can be written as $C_{\boldsymbol{\phi}(G_{i-1})}(F_{1,\gamma_1}(.), ..., F_{d,\gamma_d}(.))$, and the conditional distribution function $H_{G_{i-1}}$ of $\mathbf{X}_i$ given $G_{i-1}$ can be written as:

$$H_{G_{i-1}}(\mathbf{x}) = C_{\boldsymbol{\phi}(G_{i-1})}\left(F_{1,\gamma_1}\Big(\frac{x_1 - \mu_{i1}(\boldsymbol{\beta}_1)}{\sigma_{i1}(\boldsymbol{\beta}_1)}\Big), ..., F_{d,\gamma_d}\Big(\frac{x_d - \mu_{id}(\boldsymbol{\beta}_d)}{\sigma_{id}(\boldsymbol{\beta}_d)}\Big)\right), \mathbf{x} \in R^d. \tag{3.6}$$

If $c_\theta$ is the density of $C_\theta$ and $f_{j,\gamma_j}$ is the density of $F_{j,\gamma_j}$, then the conditional density $h_{G_{i-1}}$ of $\mathbf{X}_i$ given $G_{i-1}$ is:

$$\begin{aligned} h_{G_{i-1}}(\mathbf{x}) =& c_{\boldsymbol{\phi}(G_{i-1})}\left(F_{1,\gamma_1}\Big(\frac{x_1 - \mu_{i1}(\boldsymbol{\beta}_1)}{\sigma_{i1}(\boldsymbol{\beta}_1)}\Big), ..., F_{d,\gamma_d}\Big(\frac{x_d - \mu_{id}(\boldsymbol{\beta}_d)}{\sigma_{id}(\boldsymbol{\beta}_d)}\Big)\right) \cdot \\ & \cdot \prod_{j=1}^{d} \frac{1}{\sigma_{ij}(\boldsymbol{\beta}_j)} f_{j,\gamma_j}\Big(\frac{x_j - \mu_{ij}(\boldsymbol{\beta}_j)}{\sigma_{ij}(\boldsymbol{\beta}_j)}\Big). \end{aligned} \tag{3.7}$$

It is possible to estimate the model through the log-likelihood:

$$l_n(\boldsymbol{\beta}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_d, \gamma_1, ..., \gamma_d) = log \prod_{i=1}^{n} h_{G_{i-1}}(\mathbf{X}_i) = \sum_{i=1}^{n} log\, h_{G_{i-1}}(\mathbf{X}_i). \tag{3.8}$$

The d components of the time series are estimated separately, if $\boldsymbol{\beta}_{j,n}$ and $\gamma_{j,n}$ are the estimators of $\boldsymbol{\beta}_j$ and $\gamma_j$, then the standardized residuals are estimated through:

$$\epsilon_{ij,n} = \frac{X_{ij} - \mu_{ij}(\boldsymbol{\beta}_{j,n})}{\sigma_{ij}(\boldsymbol{\beta}_{j,n})}. \tag{3.9}$$

The parameter $\boldsymbol{\beta}$ of $\phi$ is estimated by:

$$argsup_\beta \sum_{i}^{n} log\, c_{\boldsymbol{\phi}(G1_{i-1})}(F_{1,\gamma_{1,n}}(\epsilon_{i1,n}), ..., F_{d,\gamma_{d,n}}(\epsilon_{id,n})). \tag{3.10}$$

If the conditional copula $G_{i-1}$ does not depend on $G_{i-1}$, then $\phi = \theta$ and the parameter $\boldsymbol{\beta}$ is the same as the parameter $\theta$ of the copula and $\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_d$ are individually identically distributed. The estimator $\boldsymbol{\beta}_n$ of $\boldsymbol{\beta} = \theta$ is:

$$\boldsymbol{\beta}_n = argsup_{\theta \in \Theta} \sum_{i}^{n} log\, c_\theta(F_{1,n}(\epsilon_{i1,n}), ..., F_{d,n}(\epsilon_{id,n})), \tag{3.11}$$

where the univariates are nonparametrically estimated with the empirical distribution function:

$$F_{j,n}(x) = \frac{1}{n+1} \sum_{i=1}^{n} 1(\epsilon_{ij,n} \leq x), \quad x \in R, \tag{3.12}$$

so $\boldsymbol{\beta}_n$ is the maximum -likelihood estimator 2.34 computed from the estimated standardized residuals 3.9.
In this setting, if the univariate time series are estimated separately instead of being jointly estimated, then the empirical copula process behaves as if the innovations were observed. As a by-product, one also obtains the asymptotic behavior of rank-based measures of dependence applied to residuals of these time series models [11].
This means that it is possible to apply the the rank-based inference method presented in the previous chapter on the estimated standardized residuals as if they were the innovations.

## 3.3 Clustering

Clustering the components of a multivariate time series allows to focus on specific properties of the series, while also allowing to perform a dimensional reduction by creating dimensionally smaller multivariate time series.
Defining a proper distance for time series can be problematic, for this reason copula methods are used, moreover since copulas capture the rank-invariant dependence structure of a random vector, these methods are invariant under strictly increasing transformations of the time series of interest [12].

### 3.3.1 Hierarchical clustering

As explained in [13], clustering consists in grouping objects so that objects belonging to a group are more similar to each other than objects in other groups. Hierarchical clustering is based on the concept of dissimilarity (or distance) between clusters.

Starting from the n*d matrix $X_{ij}$ of the observations, the objective is to group the columns into K clusters, the division is based on the degree of similarity between objects, this information is provided by a dissimilarity matrix $\Delta_{ij}$ of size d*d, for this matrix the following properties hold:

1. $\Delta_{ij} \geq 0$ for every $i, j \in \{1, ..., d\}$.
2. $\Delta_{ij} = \Delta_{ji}$ for every $i, j \in \{1, ..., d\}$.
3. $\Delta_{ii} = 0$ for every $i \in \{1, ..., d\}$.

So $\Delta_{ij}$ is the degree of dissimilarity between the i-th and the j-th column of the data matrix, and it is low for similar objects and high for dissimilar objects. In hierarchical clustering the clusters are created either in an agglomerative or a divisive way. In the agglomerative approach each column start as its own cluster and then the closest clusters are gradually merged together until all columns are in a single cluster or until a stopping criterion is met; in the divisive approach all columns starts in the same cluster and then they are recursively split into smaller subclusters, until all columns are in their own cluster or until a stopping criterion is met.
The dissimilarity between clusters can be computed in various ways, some possible choices could be the maximum dissimilarity between the elements of the clusters (complete method), or the average dissimilarity between the elements of the clusters (average method).
In hierarchical clustering

### 3.3.2 Copula-based hierarchical clustering

As explained so far in this chapter, once the marginals of a multivariate time series are fitted and their residuals are extracted, it is possible to obtain the matrix $X_{ij}$ of pseudo-observations, where each column correspond to a component of the time series, and study the dependencies between the components of the time series via copulas.
In copula based clustering the dissimilarity is based on the properties of the bivariate copulas $C(\mathbf{X}_i, \mathbf{X}_j)$ associated to the columns i and j.
As previously explained in sections 2.2.1 and 2.2.2 it is possible to estimate nonparametrically, so without making any assumption on the underlying copula, the rank correlation measures and the tail dependence coefficient, and hence obtain a matrix $\hat{M}_{ij}$ where the element $\hat{m}_{ij}$ is the estimated coefficient for the pair of columns i and j. The definition of the dissimilarity matrix changes based on the coefficient considered, typically for Spearman's rho and Kendall's tau it is defined as $\Delta_{ij} = \sqrt{1 - \hat{M}_{ij}^2}$, while a common choice

for the tail dependence coefficient is $\Delta_{ij} = -log(\hat{M}_{ij})$ [14], but as long as the three properties of the dissimilarity matrix are respected other choices are possible.
The choice of the coefficient to consider is based on the aim of the clustering, spearman's rho and kendall's tau are used when the interest is in the general association of the components, while the tail dependence coefficient is used when the interest is in the tail behaviour and the co-occurrences of extreme events.

# Chapter 4

# Data analysis

The data used is freely available on the ARPAFVG website [15]. The original dataset is composed of daily weather data of multiple weather stations located in the Friuli-Venezia Giulia region.
For the analysis done in this thesis only the stations with 20 years of data, starting from January 2004 up to December 2023, were considered; moreover for each station only the time series made of the monthly maximum daily rainfall[mm] is considered.
The analysis was done in R with the help of the copula package [16].
In figure 4.1 and in table 4.1 it is possible to see a recap about the 18 weather station used for the analysis. It is noticeable that the weather stations are

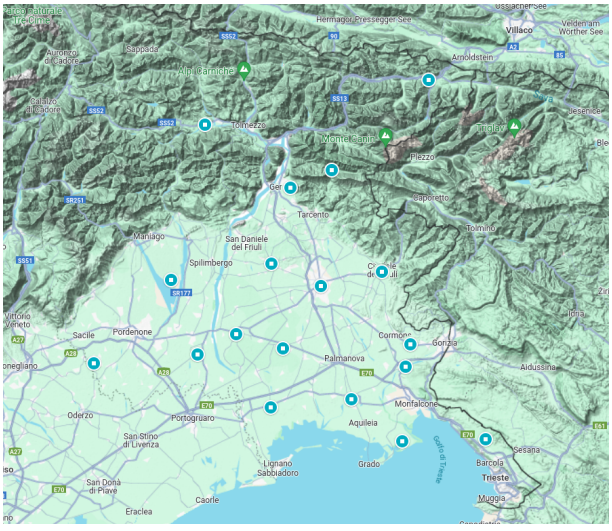


Figure 4.1: Map of used weather stations.

concentrated in the central part of the region, and that in the northern part there are fewer stations and they are more sparse, this is due to the unavailability of at least twenty years of continuous data for most of the station of that area.

| ***Locality*** | ***Altitude A.S.L.*** $[m]$ | ***Latitude*** | ***Longitude*** |
|---|---|---|---|
| Brugnera | 22 | 45.91792 | 12.54500 |
| Capriva del Friuli | 85 | 45.95809 | 13.51233 |
| Cervignano del Friuli | 8 | 45.84949 | 13.33701 |
| Cividale del Friuli | 127 | 46.08044 | 13.42001 |
| Codroipo | 37 | 45.95236 | 13.00274 |
| Enemonzo | 438 | 46.41042 | 12.86254 |
| Fagagna | 148 | 46.10169 | 13.07389 |
| Fossalon | 0 | 45.71477 | 13.45886 |
| Gemona del Friuli | 184 | 46.26130 | 13.12209 |
| Gradisca d'Isonzo | 29 | 45.88979 | 13.48181 |
| Musi | 600 | 46.31266 | 13.27468 |
| Palazzolo dello Stella | 5 | 45.80572 | 13.05260 |
| San Vito al Tagliamento | 21 | 45.89566 | 12.81499 |
| Sgonico | 268 | 45.73800 | 13.74206 |
| Talmassons | 16 | 45.88231 | 13.15779 |
| Tarvisio | 794 | 46.51078 | 13.55189 |
| Udine | 91 | 46.03521 | 13.22667 |
| Vivaro | 142 | 46.07653 | 12.76881 |

Table 4.1: Information about used weather stations.

## 4.1 Data exploration

In figure 4.2 it is possible to see the univariate time series, the time series appear to be stationary, this fact was also confirmed by running the augmented Dickey-Fuller test. To check wether they are also individually identically distributed (iid) the auto correlation plots were studied. In figure 4.3 it is possible to see the autocorrelation plots, they display the correlation between the time series and a lagged version of themselves, a low autocorrelation indicates that the time series can be considered iid. From the plots it can be seen that most of the time series are iid, but there are four stations that exhibit a problematic pattern, they are: Enemonzo, Gemona del Friuli, Musi, and Tarvisio. The pattern in the plots suggest a seasonality effect, this can

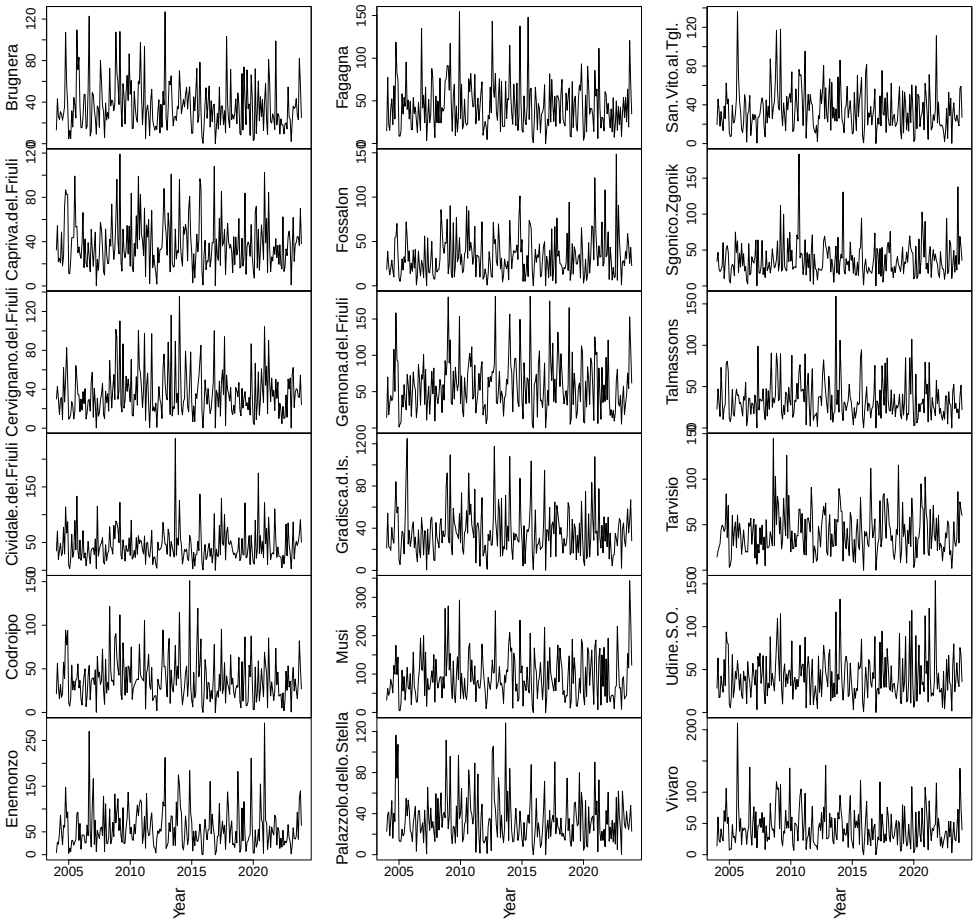


Figure 4.2: Plots of the univariate time series.

be explained by the fact that, as it can be seen from table 4.1 and figure 4.1, these four stations are situated in the northern mountainous area of the region, so in the winter months it is more likely that it will snow. To confirm the iid hypothesis the Ljung-Box test was run, the p-value of the test confirms what previously said, only the four aforementioned stations have a low p-value, so these time series need to be treated differently to account for the autocorrelation.

## 4.2 Marginal modeling

A general suitable model for the univariate time series is the autoregressive integrated moving average (ARIMA). In the ARIMA(p,d,q) model the

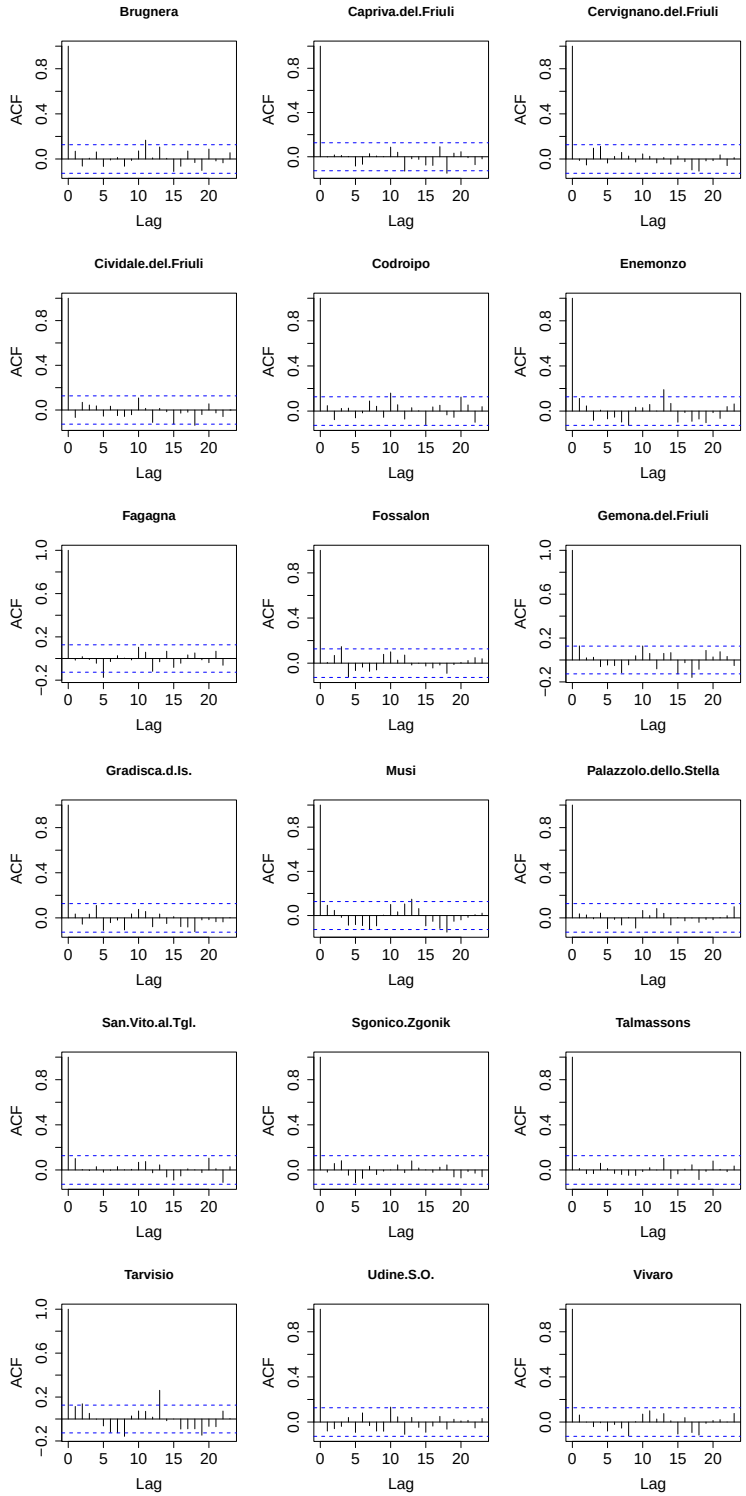


Figure 4.3: Plots of the ACF of the univariate time series.

parameter d indicates the degree of differencing (necessary in the case the time series is not already stationary), the parameter p and q are respectively the degree of the AR and MA models. The general ARMA(p,q) has the form:

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}, \tag{4.1}$$

where $\phi_0 = \mu(1 - \phi_1 - ... - \phi_p)$ is the intercept of the model and $(\epsilon_t)_{t \in Z}$ is the process of innovations.
The residuals are inferred values $\hat{\epsilon_t}$ for the unobserved errors $\epsilon_t$. Since all the time series are already stationary the models considered were ARIMA(p,d=0,q) which is equivalent to the ARMA(p,q).
For the time series that are both stationary and iid, the best model to fit them is the ARMA(0,0). But, since the time series are iid, there is no need to fit the data and compute the residuals, because these observations can be transformed directly to pseudo-observations.
For the four time series that are not iid a suitable order for the ARMA model was chosen and the residuals of the fitted model were transformed to pseudo-observations. As a way to choose the most appropriate order the R function auto.arima from the forecast package [17] was used, the function was run on all time series and it confirmed that the order d of the ARIMA is 0 for each time series, as well as the choice of the ARMA(0,0) for the iid series, while for the remaining series the orders found can be seen in the table 4.2. Having

| **Station** | **p** | **q** |
|---|---|---|
| Enemonzo | 1 | 0 |
| Gemona del Friuli | 1 | 0 |
| Musi | 2 | 2 |
| Tarivsio | 1 | 1 |

Table 4.2: ARMA(p,q) orders for the non iid time series.

fitted the models and having computed the residuals of these four time series, it was them possible to obtain the pseudo-observations for the whole data. Starting from the data matrix $X_{ij}$, with $i \in \{1, ..., n\}$ number of observations and $j \in \{1, ..., d\}$ number of weather stations, i.e. each colums is made of the observations in the case of the iid time series, and of the residuals in the case of the four fitted time series, the pseudo-observations matrix was obtained as:

$$U_{ij} = \frac{R_{ij}}{n+1},$$

where $R_{ij}$ the rank of $X_{ij}$ among $X_{1j}, ..., X_{nj}$.

## 4.3 Copula fit

As previously stated in section 2.4.2 if the dimension d of the copula is greater than two, the graphical exploration and the statistical test can be done on all the $\binom{d}{2}$ different bivariate margins.

### 4.3.1 Graphic exploration

Using the pseudo-observations just computed, a first exploratory analysis was done by plotting the pairwise scatter plots matrix, which can be seen in figure 4.4,

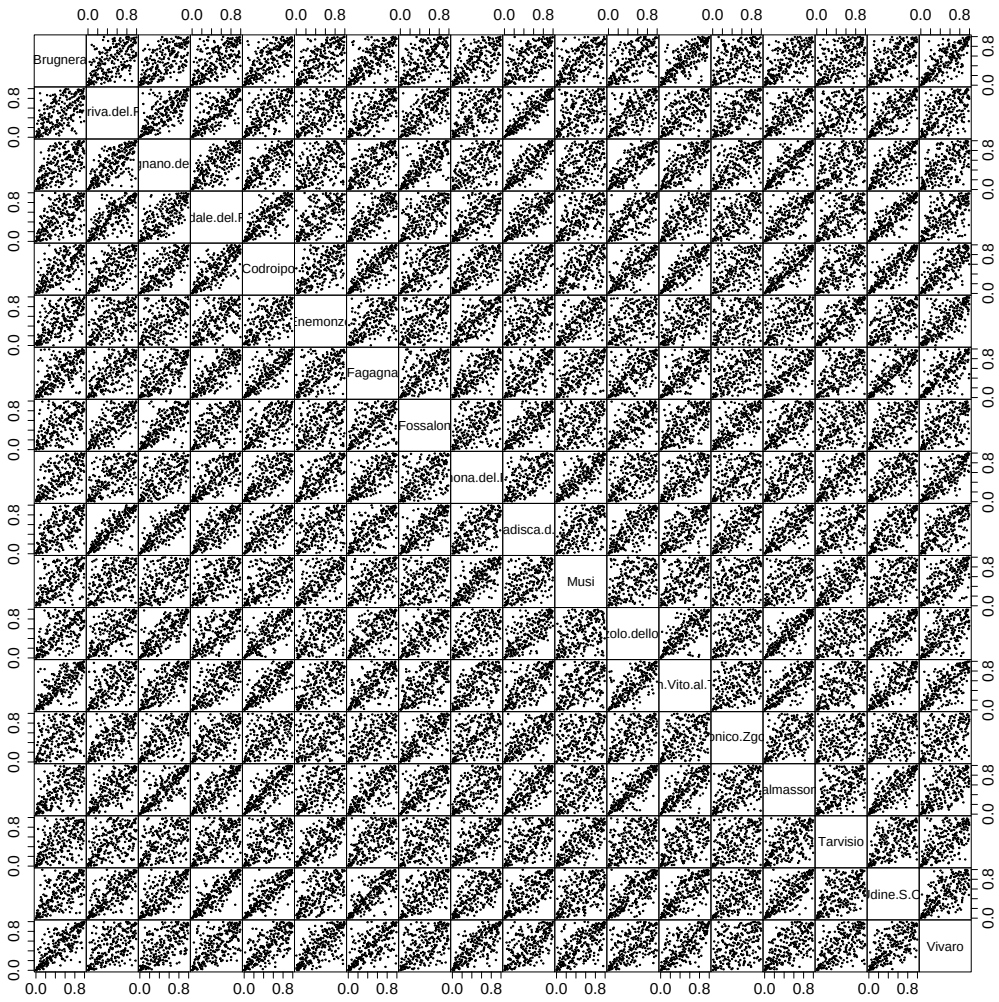


Figure 4.4: Pairwise scatter plots matrix of all stations pairs.

from the scatter plots it appears that there is a positive dependence between all the pairs and that its strength could depend on the geographical distance, as it can be seen from the Sgonico row/column, which is the farthest station with respect to all the others; the same relationship seems to hold also for the upper tail dependence coefficient, while it appears that the lower tail dependence is strong for all pairs, except for the Enemonzo row/column for which the lower tail dependence is smaller but still not zero.
These hypothesis are reinforced by the estimations of the Kendall tau and of the tail dependence coefficients, a heatmap of these values can be seen in the figures 4.5 and 4.6, in these graphs a lighter colour indicates a weaker dependence, so it is easier to notice what hypothesized above.

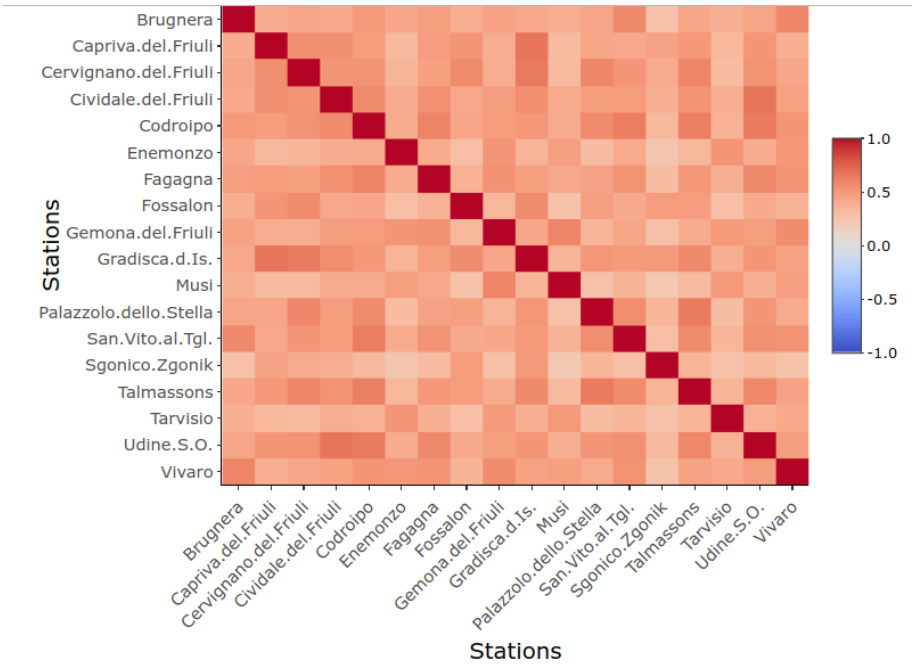


Figure 4.5: Heatmap of the estimated Kendall tau.

Another method used to study the bivariate dependencies is the K-plot, which consists in the plot of $(W_{i:n}, H_{(i)})$ for $i \in \{1, ..., n\}$, where $H_{(i)}$ are the order statistics associated to the quantities:

$$H_i = \frac{1}{n-1} \#\{j \neq i : X_{1j} \leq X_{1i}, X_{2j} \leq X_{2i}\}$$

and $W_{i:n}$ is the expected value of the i-th statistic from a random sample of size n from the random variable $W = C(U, V) = H(X_{1i}, X_{2j})$ under the

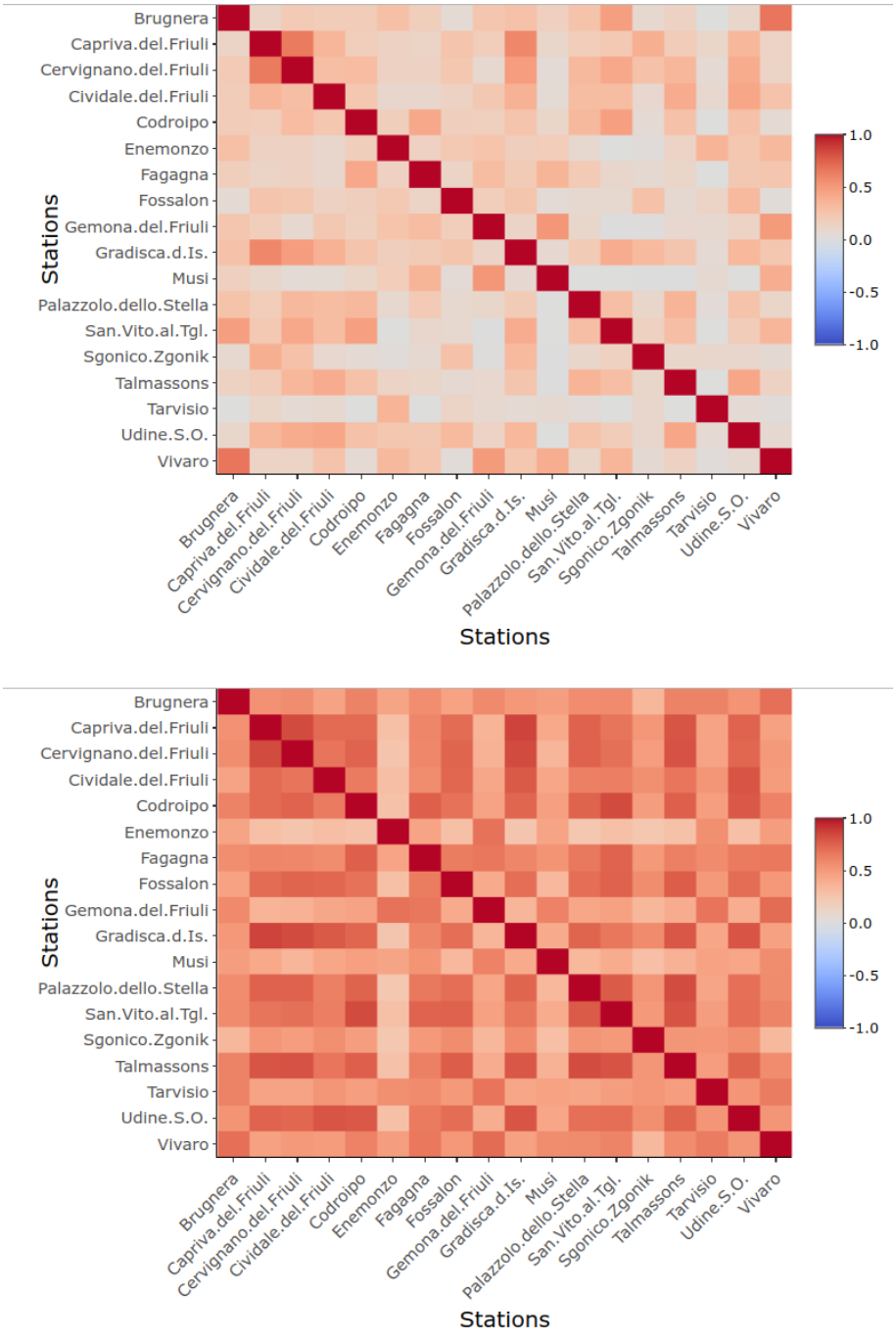


Figure 4.6: Heatmap of the estimated upper (above) and lower (below) tail dependence coefficients.

null hypothesis of independence between U and V (or between $X_{1i}$ and $X_{2j}$, which is the same) [9].
The plot for all pairs was studied, but since plotting the function for all available pairs is not practical, three significant examples can be seen in figure 4.7, The diagonal line is the case of independence and the curve is the case of positive perfect dependence, in the plots it can be seen that in all three cases there is a strong dependence in the lower tail, and that in the plots from left to right the strength of the dependence and the upper tail dependence coefficient both increase.
What seen from the other K-plots confirmed the aforementioned hypothesis.

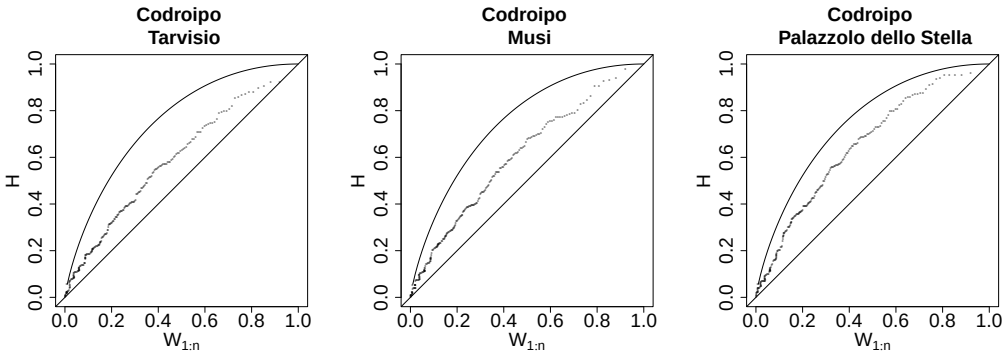


Figure 4.7: Example of K-plots for the pairs, from left to right, Codroipo/Tarvisio, Codroipo/Musi, Codroipo/Palazzolo dello Stella.

### 4.3.2 Statistical tests

To test other hypothesis some statistical test were run, at first the exchangeability was tested, if the pair of station $X_{ij}$ can be considered exchangeable then the pair $X_{ji}$ can be ignored, allowing for a first dimensionality reduction.
The test returned a low p-value only for a handful of pairs, like for example the pair Musi/San Vito al Tagliamento, but looking at the scatter plot of the pair in figure 4.4 the non-exchangeability does not appear to be in the tails and is probably caused by the low amount of data, so in the analysis these pairs are considered exchangeable anyway. So essentially all pairs can be considered exchangeable allowing to half the numbers of pairs to consider, i.e. only the lower half of the matrix of figure 4.4.

Other tests that were run are the radial symmetry test, which signaled around half of the pairs as radially symmetric, and the extreme-value dependency test, but none of the pairs had a p-value over the threshold, so the hypothesis of extreme-value dependence was rejected, meaning that the eventual copulas that will be fitted won't belong to the extreme-value family.

### 4.3.3 Clustering

Since there are $d = 18$ weather stations, fitting a d-dimensional copula and fitting the $\frac{1}{2}d(d-1)$ pairwise copula is too computationally expensive, so a farther dimensionality reduction was required, to do so the clustering method previously described was implemented.
Given what was found in the exploratory analysis, the estimated upper tail dependence is a good candidate for building the dissimilarity matrix which, in this case, is computed as:

$$\Delta_{ij} = \sqrt{1 - \hat{\lambda}_{ij}}, \tag{4.2}$$

where $\hat{\lambda}_{ij}$ is the estimated upper tail dependence coefficient for the couple of stations i and j.
In order to select the clusters the dendrogram is considered, in this plot the y axis represent the dissimilarity at which the clusters merge, while on the x axis the elements of the clusters are placed. In figure 4.8 it is possible to see the dendrograms that was obtained, the horizontal blue line is the height level at which the clusters are created, using the average method three clusters would be created , with one of them containing more than half of the stations; with the complete method five clusters would be created, with the biggest containing six stations, the following analysis will be done considering the clusters found with the complete method, a recap of the clusters can be found in the table 4.3.
The pairs Tarvisio/Enemonzo and Brugnera/Vivaro each form a single cluster because, as it can be seen from the upper tail dependence in the plot in figure 4.6, Tarvisio has a significant tail dependence only with Enemonzo, and the tail dependence between Brugnera and Vivaro is quite bigger than the one of the other pairs formed by these two stations, so they have a higher dissimilarity, as it can be seen from the dendrogram in figure 4.8.

The clusters contain the stations for which the co-occurences of extreme rainfall are more likely, looking at the map in figure 4.9 there appears to be a geographical component to the clusters, as close stations belong to the same

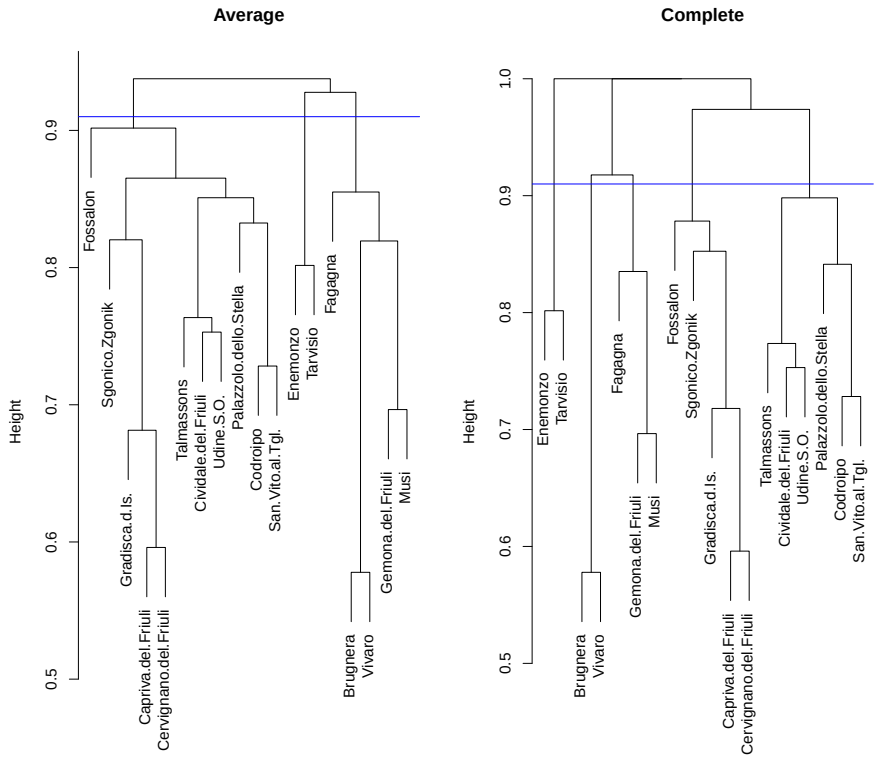


Figure 4.8: Dendrograms of the hierarchical clustering, produced with the average (left) and complete (right) methods.

cluster.
To check the internal validity of the clusters, for each of them the estimated upper tail coefficient heatmap, the K-plots and the pairwise scatter plots were studied, it is expected that this plots behave similarly for all the pairs of elements of a cluster.
In the following only what was done on the cluster 4 is explained, the work done on the other clusters being analogous.
In the figure 4.10 the pairwise scatter plots matrix of the pseudo-observations can be seen, from these plots it looks like the upper tail dependence coefficient does not change too much between pairs, it can be noticed that the same seems to be true also for the lower tail dependence coefficient and for the dependence outside the tails.
The upper tail dependence coefficient heatmap can be seen in figure 4.11,

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|
| Enemonzo | Brugnera | Talmassons | Fagagna | Fossalon |
| Tarvisio | Vivaro | Gemona del Friuli | Cividale del Friuli | Sgonico |
| | | Musi | Udine S.O. | Gradisca d'Isonzo |
| | | | Palazzolo dello Stella | Capriva del Friuli |
| | | | San Vito al Tagliamento | Cervignano del Friuli |
| | | | Codroipo | |

Table 4.3: Clusters found with the complete method.

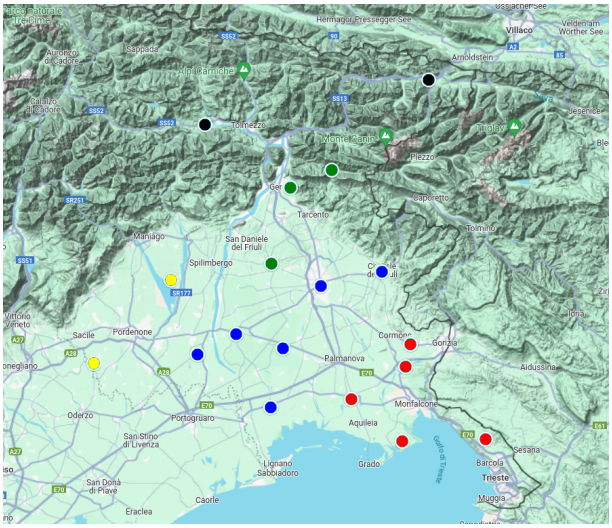


Figure 4.9: Map of the clusters; cluster 1:black, cluster 2:yellow, cluster 3:green, cluster 4:blue, cluster 5:red.

as expected this confirms the fact that the coefficient is similar for all the pairs. Finally in the figure 4.12 the pairwise K-plots can be seen, looking at the upper right and the lower left part of the plots the upper and lower tail dependence seems to be similar for all pairs, moreover looking at the middle part of the plots it looks like the dependencies between pairs outside the tails is also similar, as it was also noticed from the scatter plots.

### 4.3.4 Estimation and Goodness of fit

Having built the clusters, it was then possible to estimate the copulas that best fit them, for each cluster a d-dimensional copula, where d is the number of stations in the cluster, was fitted.
The process followed was the same for each cluster, so in the following only what was done for cluster 4 is presented, the others being analogous.
In order to choose which copula family to fit at first the hypothesis of radial

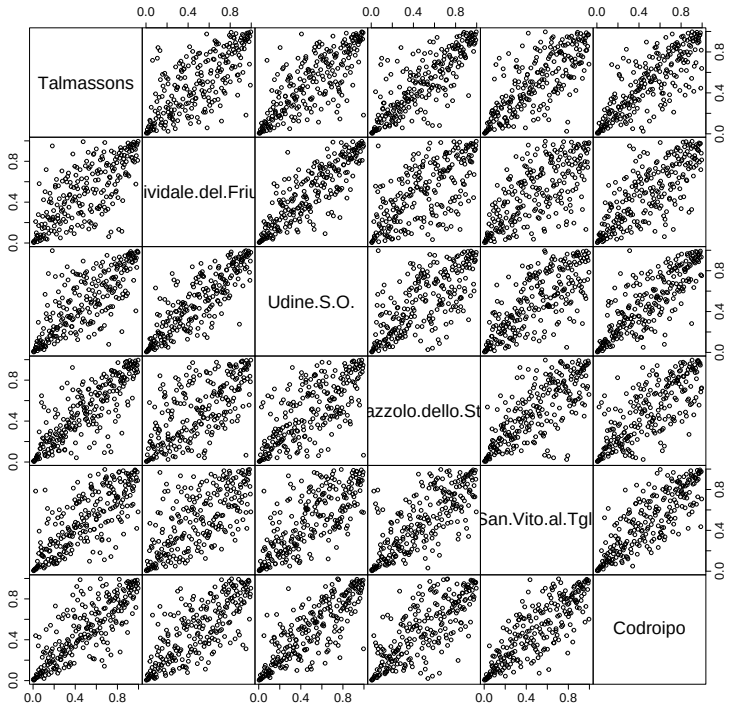


Figure 4.10: Pairwise scatter plots of stations in cluster 4.

symmetry and exchangeability were tested, in the case of cluster 4 both hypothesis were confirmed, so it is likely that the eventual copula belongs to the elliptical family. Looking at the pairwise scatter plots in figure 4.10 and at the upper tail dependence coefficient heatmap in figure 4.11, the big tail dependencies suggest that a t copula, rather than a normal copula would be a good fit.

Both the t and normal copula were fitted, the p-value of the goodness of fit test for the both copula was over threshold, but the t copula was chosen as the best fit according to the result of the AIC test, the AIC was also computed for other copula families but their AIC was too big to be considered. All the copula were estimated using the maximum pseudo likelihood method 2.34, and tested using the test statistic 2.44.

Doing the same for the other clusters it was possible to fit a copula on all the clusters, a recap of the best fitting copula can be found in table 4.4, for all the clusters the best fitting copula belongs to the elliptical family, in particular for the first cluster the best one is the normal copula, while for the others it is the t copula.

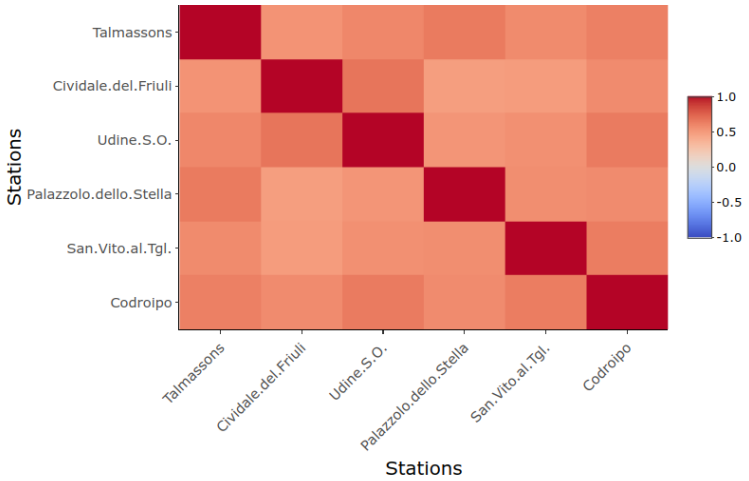


Figure 4.11: Heatmap of the estimated upper tail dependence coefficients of stations in cluster 4.

## 4.4 Findings

Thanks to copulas it was possible to identify five different subregions of Friuli-Venezia Giulia for which the co-occurrence of extreme rainfall is more likely; as previously stated this subregions present a strong spatial division despite the fact that no spatial information was used in the analysis.
The spatial division of clusters probably also influenced the choice of the best fitting copula, indeed the clusters are formed based on the upper tail dependence, and given the geographical vicinity of the elements of a cluster the lower tail dependence is also high, from this follows a radial symmetry in addition to the exchangeability, the presence of both symmetries is a property of the elliptical copula.

| | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** | **Cluster 5** |
|---|---|---|---|---|---|
| **Dimension** | 2 | 2 | 3 | 6 | 5 |
| **Type** | Normal copula | t copula | t copula | t copula | t copula |
| **Dispersion structure** | Exchangeable | Exchangeable | Unstructured | Unstructured | Unstructured |
| **Parameters** | $\rho = 0.733$ | $\rho = 0.808$ | $\rho_{avg} = 0.735$ | $\rho_{avg} = 0.786$ | $\rho_{avg} = 0.750$ |
| | | | $\rho_{min} = 0.642$ | $\rho_{min} = 0.660$ | $\rho_{min} = 0.690$ |
| | | | $\rho_{max} = 0.815$ | $\rho_{max} = 0.860$ | $\rho_{max} = 0.875$ |
| | | $df = 3.00$ | $df = 8.00$ | $df = 5.00$ | $df = 5.00$ |

Table 4.4: Best fitting copula family.

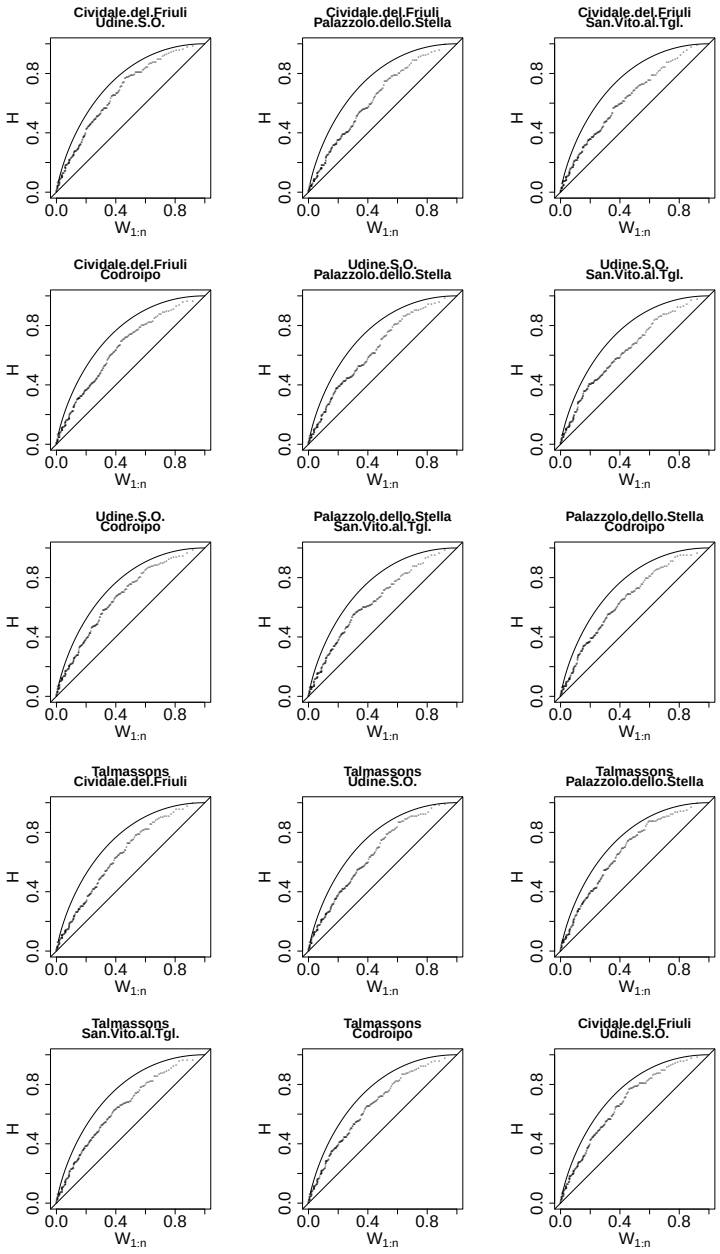


Figure 4.12: Pairwise K-plots of stations in cluster 4.

# Chapter 5

# Conclusions

The thesis discusses the use of copula functions in variable clustering and present an application to climate time series.
In particular, the thesis focused on a time series clustering procedure suitable for identifying groups of meteorological stations dependent on the upper tail. The use of copula allowed to model the individual time series of monthly rainfall maxima separately from each other and then study their dependencies through the model residuals.
At first all the tail dependence coefficients, which describe the dependence in the tails of the joint distribution via the copula of the data, were estimated. These estimations were then used to define a dissimilarity matrix to be used to cluster together the weather stations for which the co-occurrence of extreme rainfall is more likely.
Once a final cluster solution was identified, the result was used to explore the dependency structure within each cluster by fitting various parametric copulas. This, in turn, allowed to avoid estimating high-dimensional models, making computationally less demanding. In particular five different subregions were identified, and on each of the five clusters it was possible to fit a suitable copula, selected among various models by means of statistical tests and graphical tools.
As for possible future improvements, a larger number of stations and observations could be considered to improve the presented clustering method and enhance the quality of the results. Moreover, a further improvement could be represented by the introduction of suitable spatial constraints in the dissimilarity construction, to explicitly account for the spatial information on the observed time series in different regions.

# Bibliography


[1] A. Sklar, "Fonctions de répartition à n dimensions et leurs marges," *Publications de l'Institut de Statistique de l'Université de Paris*, vol. 8, pp. 229–231, 1959.

[2] H. Joe, *Dependence Modeling with Copulas*. Chapman and Hall/CRC, June 2014.

[3] R. Nelsen, *An Introduction to Copulas*. Springer New York, NY, 01 2006.

[4] G. Frahm, M. Junker, and R. Schmidt, "Estimating the tail-dependence coefficient: Properties and pitfalls," *Insurance: Mathematics and Economics*, vol. 37, p. 80–100, Aug. 2005.

[5] A. J. Patton, "Modelling asymmetric exchange rate dependence," *International Economic Review*, vol. 47, p. 527–556, May 2006.

[6] A. J. Patton, "A review of copula models for economic time series," *Journal of Multivariate Analysis*, vol. 110, p. 4–18, Sept. 2012.

[7] G. De Luca and P. Zuccolotto, "A tail dependence-based dissimilarity measure for financial time series clustering," *Advances in Data Analysis and Classification*, vol. 5, p. 323–340, Dec. 2011.

[8] M. Hofert, I. Kojadinovic, M. Maechler, and J. Yan, *Elements of Copula Modeling with R*. Springer Cham, 2018.

[9] C. Genest and A.-C. Favre, "Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask," *Journal of Hydrologic Engineering*, vol. 12, pp. 347–368, July 2007.

[10] A. J. Patton, "Copula–based models for financial time series," in *Handbook of Financial Time Series* (T. Mikosch, J.-P. Kreiß, R. A. Davis, and T. G. Andersen, eds.), pp. 767–785, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

[11] B. Rémillard, "Goodness-of-fit tests for copulas of multivariate time series," *Econometrics*, vol. 5, p. 13, Mar. 2017.

[12] M. Disegna, P. D'Urso, and F. Durante, "Copula-based fuzzy clustering of spatial time series," *Spatial Statistics*, vol. 21, p. 209–225, Aug. 2017.

[13] P. D'Urso, G. De Luca, V. Vitale, and P. Zuccolotto, "Tail dependence-based fuzzy clustering of financial time series," *Annals of Operations Research*, Dec. 2023.

[14] P. R. Di Lascio F.M.L., Durante F., "Copula based clustering methods," in *Copulas and Dependence Models with Applications: Contributions in Honor of Roger B. Nelsen* (J. F. S. Manuel Úbeda Flores, Enrique de Amo Artero Fabrizio Durante, ed.), pp. 49–67, Springer International Publishing, 2017.

[15] "Arpa fvg - osmer e grn." `http://www.meteo.fvg.it/`.

[16] M. Hofert, I. Kojadinovic, M. Maechler, and J. Yan, *copula: Multivariate Dependence with Copulas*, 2023. R package version 1.1-2.

[17] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: the forecast package for R," *Journal of Statistical Software*, vol. 26, no. 3, pp. 1–22, 2008.