A COPULA APPROACH FOR MODELING ENVIRONMENTAL EXTREME EVENTS

Laureando:
Davide Sancin

Relatore:
prof.ssa Roberta Pappadà

SUMMARY

- Introduction.
- Copula definition.
- Copula method for multivariate time series.
- Copula method for time series clustering.
- Application on environmental data.
- Findings and conclusion.

INTRODUCTION

- The presentation discusses the **use** of **copula** functions in **variable clustering** and **present** an **application** to **climate time series**.
- Time series **clustering procedure suitable** for **identifying groups** of **meteorological stations dependent** on the **upper tail**.
- Use of **copula allows** to **model** the **individual time series separately** from each other, and to then **study** their **dependencies through** the **model residuals**.

COPULA DEFINITION

- **Distribution** of a **d-dimensional random variable** $\mathbf{X} = (X_1, ..., X_d)$:

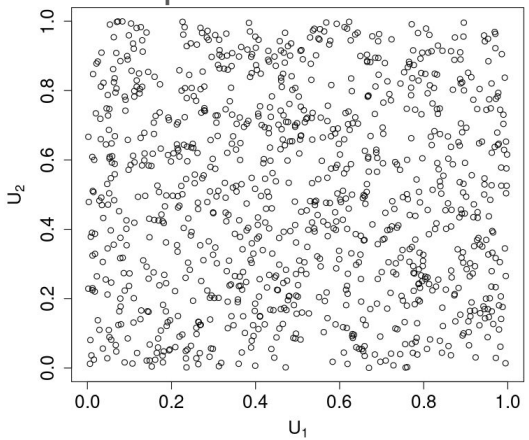
$$H(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x}) = P(X_1 \leq x_1, ..., X_d \leq x_d), \quad \mathbf{x} = (x_1, ..., x_d) \in R^d.$$

- Supposing all the **marginal distributions** F_i of $\mathbf{X}$ are **continuous** and **applying** the **probability integral transform** to each component of $\mathbf{X}$, a random vector $\mathbf{U} = (F_1(X_1), ..., F_d(X_d)) = (U_1, ..., U_d)$ with **uniform standard marginals** is obtained.
- The **copula** of $\mathbf{X}$ is then **defined** as:

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

BIVARIATE COPULA EXAMPLES

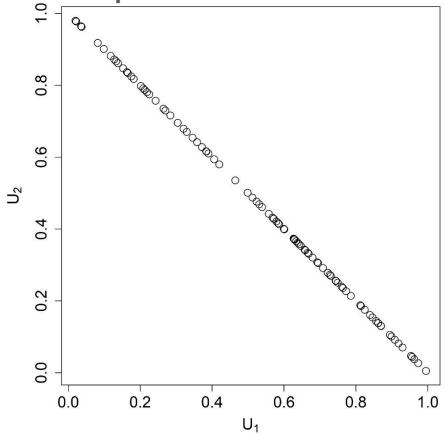
Independence



Independence copula

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

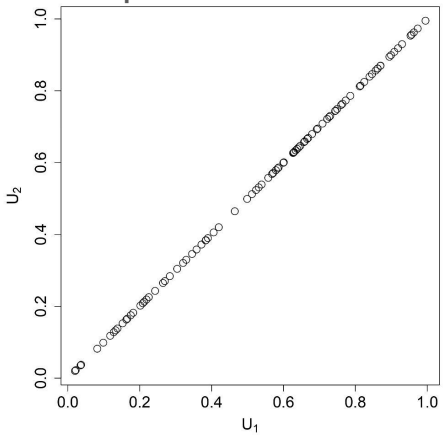
Perfect negative dependence



Countermonotone copula

$$W(\mathbf{u}) = \max\left\{\sum_{j=1}^{2} u_j - 1, 0\right\}, \quad \mathbf{u} \in [0,1]^2$$

Perfect positive dependence



Comonotone copula

$$M(\mathbf{u}) = \min_{1 \le j \le 2}\{u_j\}, \quad \mathbf{u} \in [0,1]^2$$

SKLAR'S THEOREM

- Sklar's theorem **explains** how **copulas** can **describe** the **dependencies between** the **components** of a random vector:

1. For any d-dimensional distribution function H with univariate margins $F_1,...,F_d$, there exists a d-dimensional copula C such that:

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

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

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

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

SKLAR'S THEOREM

- **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**.**
- **If** the **margins** are **continuous** then the **copula** is **unique**, **otherwise** it is **uniquely defined on** $ran\, F_1 \times ... \times ran\, F_d$ **.**
- **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**.

INVARIANCE THEOREM

- Let **X**~H with continuous univariate margins $F_1,...,F_d$ and 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**.
- **X** has copula C if and only if $(F_1(X_1), ..., F_d(X_d)) \sim C$.
- So if **X** and **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**.

TAIL DEPENDENCE COEFFICIENTS

- **Tail dependence coefficients** are used to **describe** the **dependence** in the **joint tails** of bivariate distributions.
- They are defined as **limits** of the **conditional probabilities** of **quantile exceedances**.
- The **lower** tail dependence coefficients is defined as:

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

- The **upper** tail dependence coefficients is defined as:

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

TAIL DEPENDENCE COEFFICIENTS

- Both can be **expressed exclusively** in **terms** on the **underlying copula**.
- The **lower** tail dependence:

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

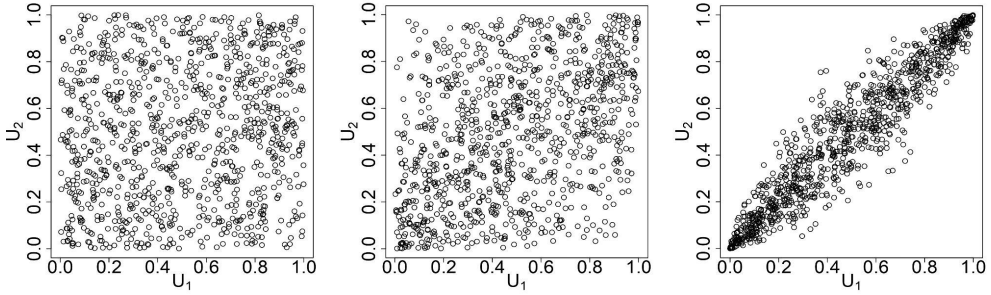
- The **upper** tail dependence:

$$\lambda_u = \lambda_u(C) = lim_{q\to 0^+} \frac{\bar{C}(q,q)}{q} = lim_{q\to 1^-} \frac{\bar{C}(1-q,1-q)}{1-q} = lim_{q\to 1^-} \frac{1-2q+C(q,q)}{1-q}$$

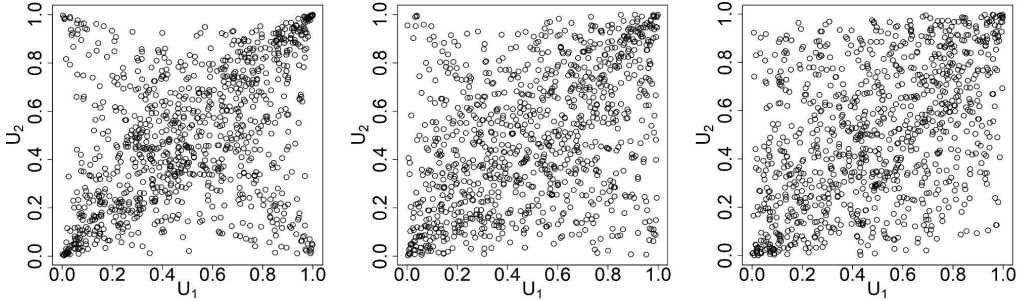
- $\bar{C}$ is the survival copula.
- Let C be a copula and let **U**~C, then 1−**U**~ $\bar{C}$
  1−**U**=(1−U_1,..., 1−U_d) is a random vector whose distribution function is the survival copula $\bar{C}$ corresponding to C.

COPULA FAMILY EXAMPLE

Normal copulas with increasing correlation coefficient



t copulas with fixed scale matrix and increasing degrees of freedom



- **Elliptical copulas** are obtained applying Sklar's theorem to elliptical distributions.
- They are both **exchangeable** and **radially symmetric**.
- Other **families** with **differing properties** exists.

COPULA-BASED TIME SERIES MODEL

- **X random vector** observed at different moments in time, the observations $(\mathbf{X}_i)_{i \in Z}$ form a **time series**.
- Each marginal have the form[1]:

$$X_{ij} = \mu_{ij}(\beta_j) + \sigma_{ij}(\beta_j)\epsilon_{ij}$$

- Possible to **model** each **univariate** time series **individually**, through classical model such as the ARIMA or the ARMA-GARCH.

[1] A. J. Patton, “Modelling asymmetric exchange rate dependence,” International Economic Review, vol. 47, p. 527–556, May 2006.

COPULA-BASED TIME SERIES MODEL

- **Residuals** $\epsilon_{i1}, ..., \epsilon_{id},\ i \in \{1, ..., n\}$ form **random sample** from an **unknown continuous distribution** with **copula** C.
- For all $j \in \{1, ..., d\}$, let R_{ij} be the rank of ϵ_{ij} among $\epsilon_{1j}, \ldots, \epsilon_{nj}$.
- $U_{ij} = F_{n,j}(\epsilon_{ij}) = R_{ij}/(n+1)$ form the **sample** of **multivariate scaled ranks**.
- From **invariance theorem**, **copula** of **multivariate scaled ranks** is the **same** as the **copula** of **residuals**, so U_{ij} is a **sample** of **pseudo-observations** from C.
- **Fit** on pseudo-observations a **proper copula** to **study** the **dependence** between the **components** of the multivariate time series.

COPULA-BASED TIME SERIES CLUSTERING

- In **environmental sciences clustering** time series can **help identifying areas** of **common risk**.
- If there are **many time series**, **fitting** a **single copula** would be **computationally demanding**, so clustering works also for **dimensionality reduction**.
- **Clustering** is **based** on the concept of **dissimilarity** (or **distance**) between clusters.
- **Defining** proper **distance** for **time series** can be **challenging**.
- **Dissimilarity** between two time series **based** on **copula measures of association**[2].

[2] 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.

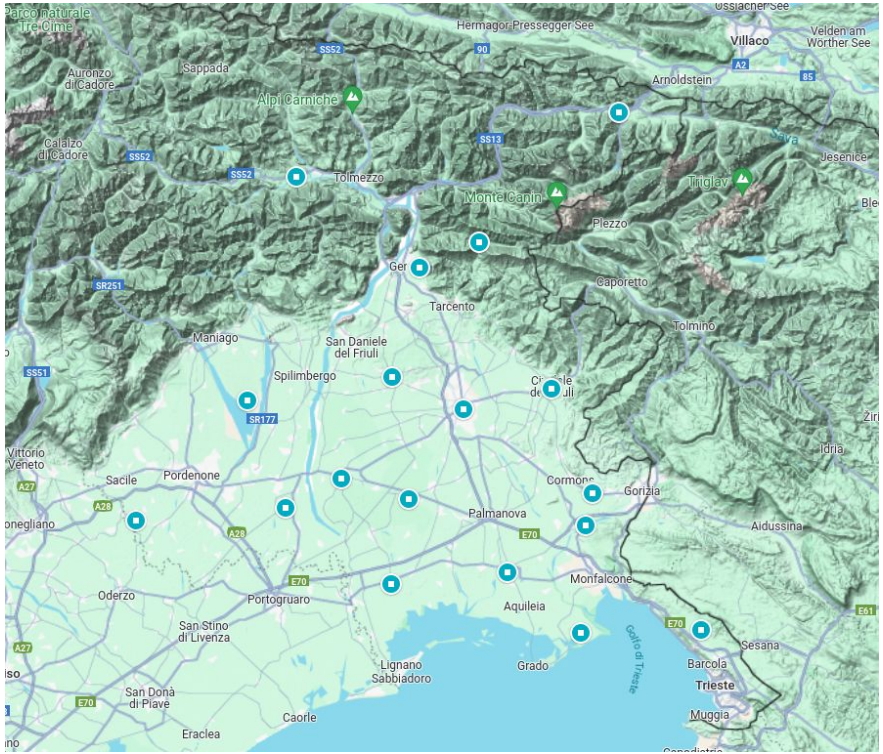
DATA EXPLORATION

- **Application** in **environmental fields** to model **extreme events**.
- Data used made of **time series** of **monthly rainfall maxima** collected at **18 weather stations** of **Friuli-Venezia Giulia**.
- Raw data available on ARPAFVG site.
- Considered only weather stations with at least **20 years of continuous data** (2004-2023).
- Analysis done in **R** with **Copula package**[3].

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

DATA EXPLORATION

Map of stations

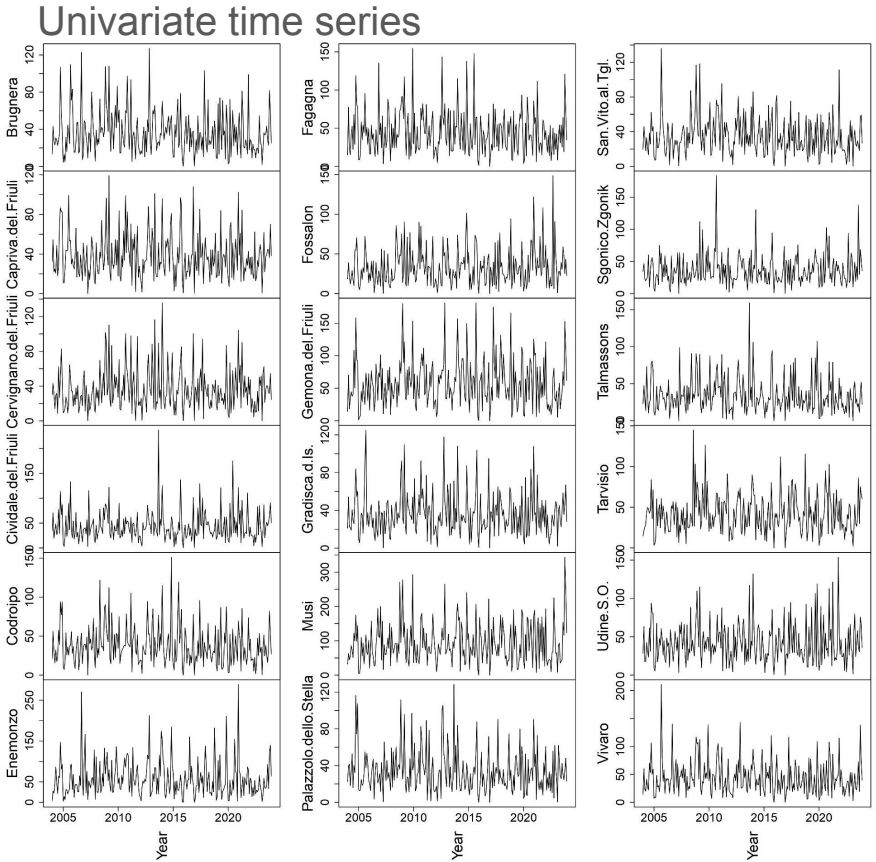


Information on stations

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

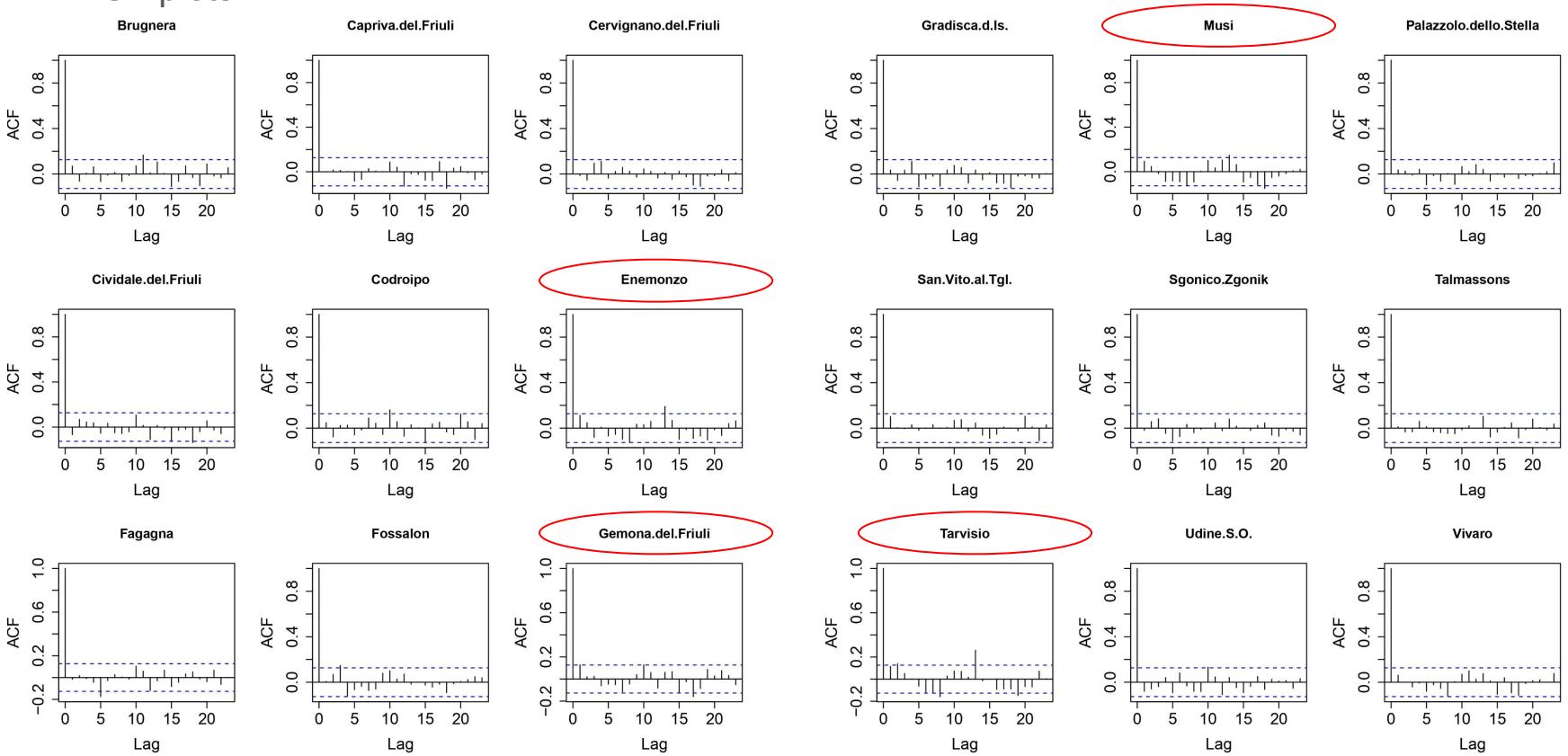
MARGINAL MODELING

- Time series appears to be **stationary**, confirmed by Dickey-Fuller test.
- To **check** whether they are also **individually identically distributed** study **autocorrelation plots**.
- They display the **correlation between** the **time series and** a **lagged version** of **themselves**.



MARGINAL MODELING

ACF plots

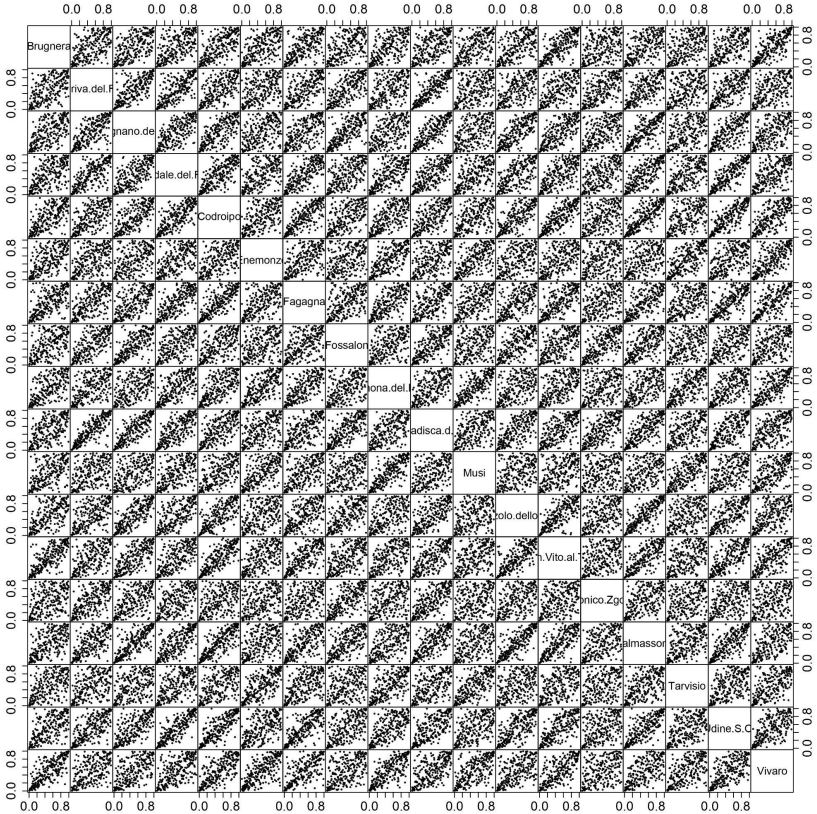


MARGINAL MODELING

- The **iid series** should be fitted with ARMA(0,0) but being iid they can be **transformed directly** to **pseudo-observations**.
- For the 4 **problematic series** a **suitable ARMA order** was selected, **then** the **residual** of the **fitted model** were **transformed** to **pseudo-observations**.
- Pseudo-observations computed as: $U_{ij} = \frac{R_{ij}}{n+1}$
- Having obtained the n*d matrix of pseudo-observations it was then possible to start the copula clustering process.

PSEUDO-OBSERVATIONS

Bivariate scatterplots matrix of pseudo-observations



- **Positive dependence** between **all** pairs.
- **Strength** appears to **depend** on **geographical distance**.
- **Same holds** for **upper tail dependence coefficient**.
- The upper tail dependence coefficient of every couple was then estimated.

TDC ESTIMATION

- **Several methods** available to estimate the **tail dependence coefficients**[4].
- **Non-parametric estimator** used for the upper tdc:

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

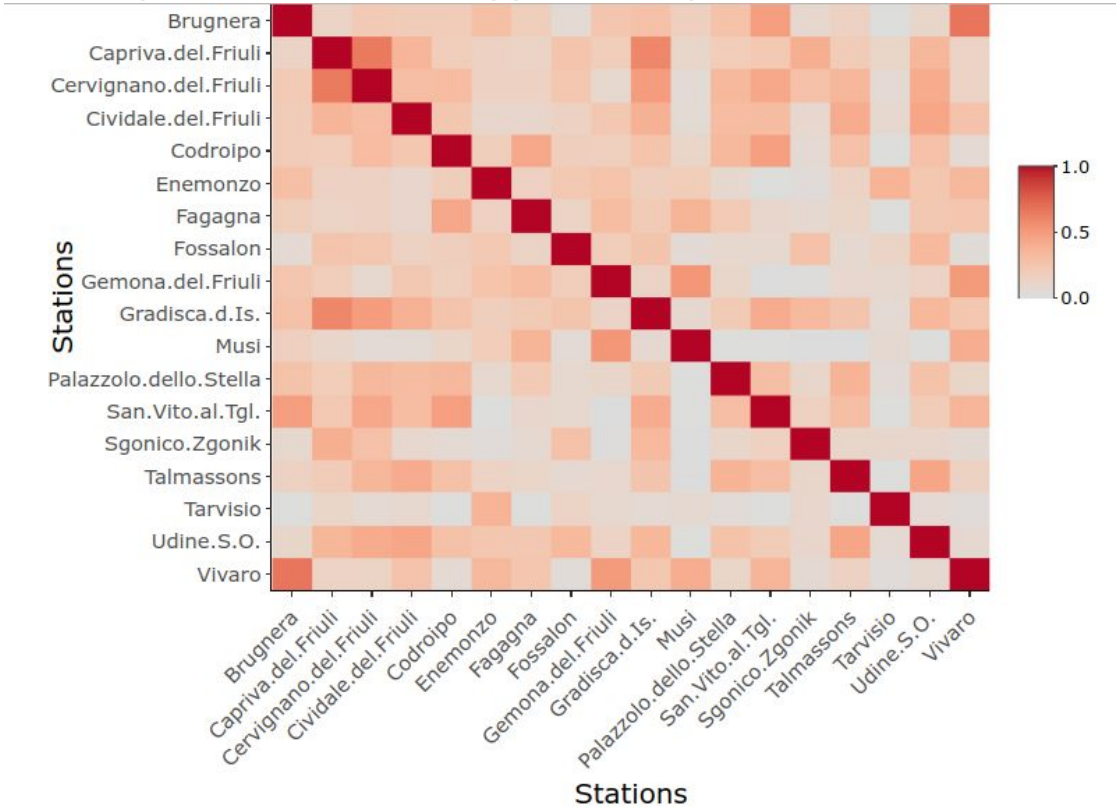
- Where the empirical copula $C_n(\mathbf{u})$ is defined as:

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

[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.

TDC ESTIMATION

Heatmap of the estimated upper tail dependence coefficients



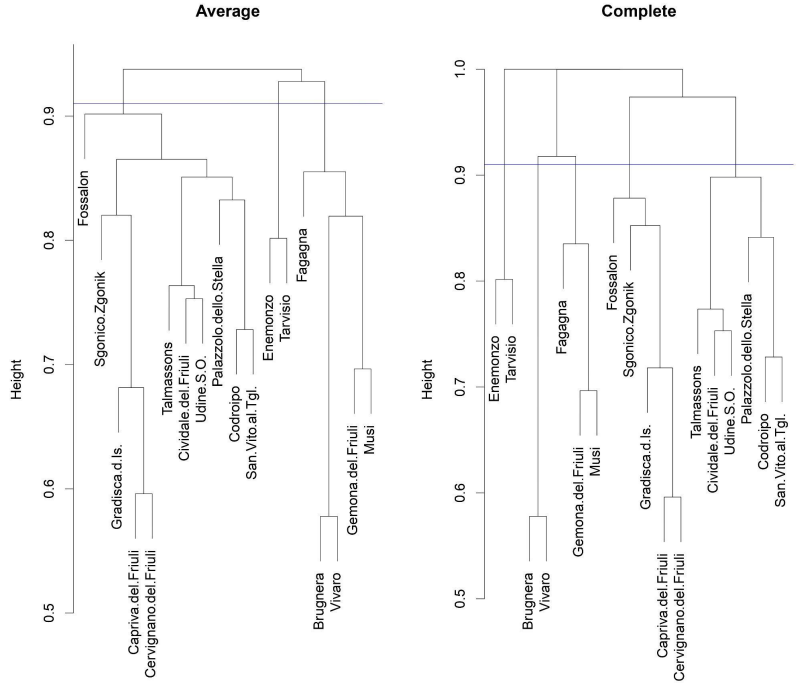
- Close to zero for most couples.
- Appears to be stronger for stations that are geographically closer.

COPULA-BASED CLUSTERING

- It was chosen to **perform** a **hierarchical clustering** based on **the upper tail dependence.**
- **Dissimilarity** used is $\Delta_{ij} = \sqrt{1 - \hat{\lambda}_{ij}}$.
- When $\hat{\lambda}_{ij}$ is one (perfect dependence) dissimilarity is minimum, viceversa when $\hat{\lambda}_{ij}$ is zero (independence) dissimilarity is maximum.
- **Stations** in **same cluster** will have **high tail dependence coefficient**, while it will be **low for stations** in **different clusters**.
- **Subregions** in which **co-occurrences** of **extreme events** are **more likely** will be **created.**

COPULA-BASED CLUSTERING

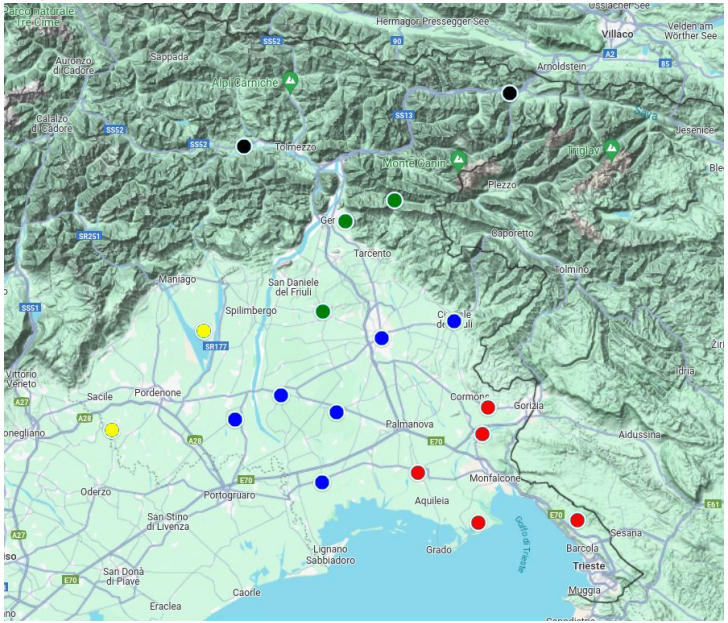
Dendrograms found with average and complete methods



- **Clusters selected** based on the **dendrogram**.
- Clusters created agglomeratively.
- **Average method** considers average dissimilarity between members of two clusters.
- **Complete method** considers maximum dissimilarity between members of two clusters.
- With average a cluster containing more than half of stations is found, complete method is better.

COPULA-BASED CLUSTERING

Map of clusters found with complete method



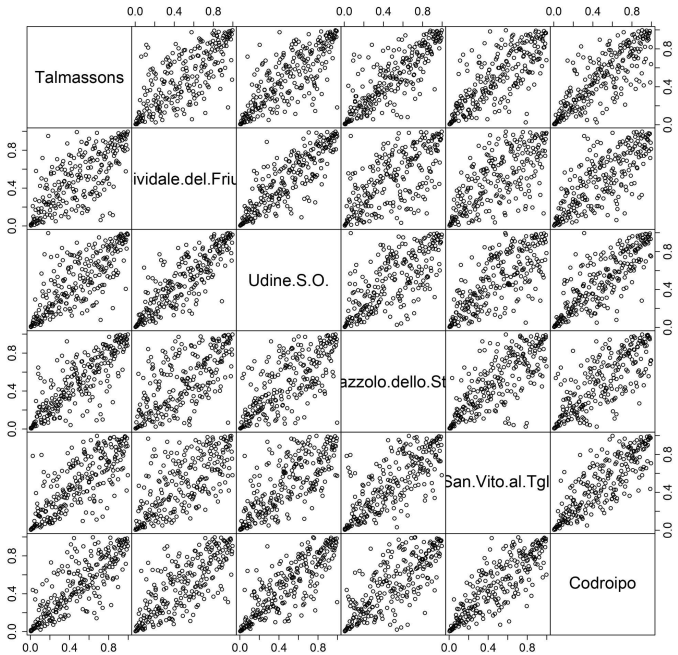
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Enemonzo	Brugnera	Talmassons	Fagagna	Fossalon
Tarvisio	Vivaro	Gemona del Friuli	Cividale del Friuli	Sgonico
		Musi	Udine S.O.	Gradisca d'Isonzo
			Palazzolo dello Stella	Capriva del Friuli
			San Vito al Tagliamento	Cervignano del Friuli
			Codroipo	

- **Five clusters** identified with the **complete method**.
- They present **spatial division**, even though **no spatial information** included.

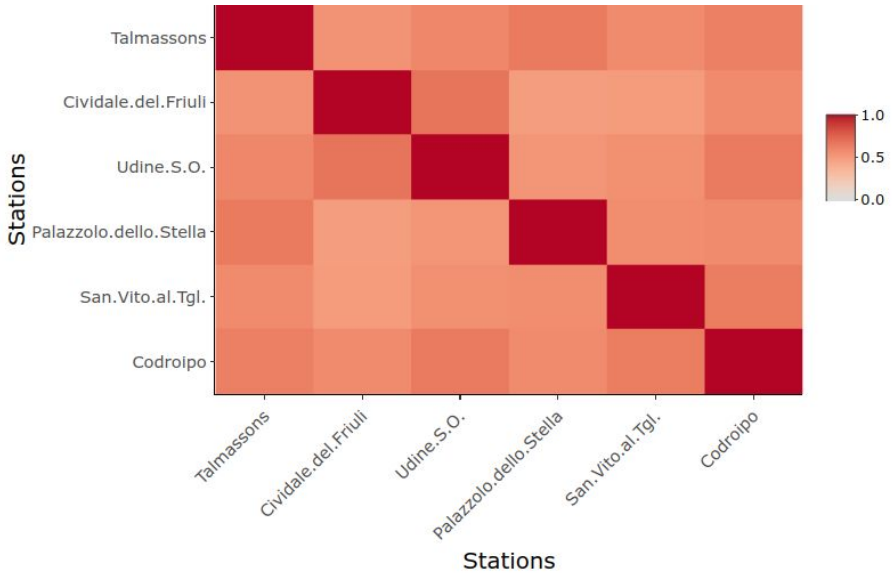
CLUSTERING VALIDATION EXAMPLE

- As a **cluster validation** the **coefficients** inside the clusters were **studied**, They should behave similarly.

Bivariate scatterplots cluster 4



Upper tail dependence heatmap cluster 4



PARAMETRIC MODEL BASED ON CLUSTERS

- Expected that **each clusters** presents a **different dependence structure**.
- On **each** of the **cluster** a **proper copula** was **fitted**.
- To **restrict** the **choice** of the **copula parametric family**, some **statistical tests** were done on the clusters.
- Notably **all clusters present** both **radial symmetry** and **exchangeability**.
- Multiple families tested**, best fitting copula** chosen based on **AIC test**.

PARAMETRIC MODEL BASED ON CLUSTERS

Best fitting copulas

	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	$\rho = 0.733$	$\rho = 0.808$	$\rho_{avg} = 0.735$	$\rho_{avg} = 0.786$	$\rho_{avg} = 0.750$
			$\rho_{min} = 0.642$	$\rho_{min} = 0.660$	$\rho_{min} = 0.690$
			$\rho_{max} = 0.815$	$\rho_{max} = 0.860$	$\rho_{max} = 0.875$
		$df = 3.00$	$df = 8.00$	$df = 5.00$	$df = 5.00$

- **All** the **copulas** are **elliptical**, expected because **both exchangeable** and **radially symmetric**.
- **Cluster 1** have **normal copula** because of the **low tail dependence** between the pair.
- All the **other** clusters have a **t copula** because of the **higher tail dependence** coefficient for the pairs.

CONCLUSIONS

- It was **possible** to **model** the **individual time series separately** from each other and **then study** their **dependencies through copulas**.
- The **tail dependence coefficients** between all pairs of stations, were **estimated**.
- These were **used** to **define** a **dissimilarity matrix** to **cluster together** the **weather stations** for which the **co-occurrence** of **extreme rainfall** is **more likely**.
- **Five different subregions** were **identified**, **on each** of the five **clusters** it was possible to **fit** a **suitable copula**.
- Include more stations and more observations to **have** a **more complete overview** of the **climate extreme events** of the region.
- **Including** a **spatial constraint** in the **dissimilarity matrix** in order to consider multiple regions.