Henri Funk, Alexander Sasse, Helmut Küchenhoff, Ralf Ludwig

Climate And Statistics

Contents

Preface

Author: Henri Funk

As the world faces the reality of climate change, natural hazards and extreme weather events have become a major concern, with devastating consequences for nature and humans. The quantification and definition of climate change, extreme events and its implications for life and health on our planet is one of the major concerns in climate science.

This book explains current statistical methods in climate science and their application. We do not aim to provide a comprehensive overview of all statistical methods in climate science, but rather to give an overview of the most important methods and their application. This book is the outcome of the seminar "Climate and Statistics" which took place in summer 2024 at the Department of Statistics, LMU Munich.

FIGURE 1: Creative Commons License

This book is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License[1].

[1] http://creativecommons.org/licenses/by-nc-sa/4.0/

Technical Setup

The book chapters are written in the Markdown language. To combine R-code and Markdown, we used rmarkdown. The book was compiled with the bookdown package. We collaborated using git and github. For details, head over to the book's repository[2].

[2] https://github.com/henrifnk/Seminar_ClimateNStatistics

1

Introduction

Author:

Supervisor:

1.1 Intro About the Seminar Topic

1.2 Outline of the Booklet

2

Introduction

Author:

Supervisor:

2.1 Intro About the Seminar Topic

2.2 Outline of the Booklet

3

Introduction

Author:

Supervisor:

3.1 Intro About the Seminar Topic

3.2 Outline of the Booklet

4

Introduction

Author:

Supervisor:

4.1 Intro About the Seminar Topic

4.2 Outline of the Booklet

5

Flood Frequency Analysis

Author: Hannes Grün, Robin Schüttpelz

Supervisor: Henri Funk

Suggested degree: Master

Abstract

5.1 Introduction

TODO: Add something "decoupling"… bla

5.2 Data

Hannes: Ist meine Variablenbeschreibung korrekt? Könntest du noch definieren, was ein baseflow und was ein streamflow ist?

? used the variables peak, volume and duration of the most severe flood event within a year. These variables are derivable from yearly hydrological discharge data. Discharge, measured in $[m^3/s]$, denotes the volume of water passed through a river within 1 second of time. The discharge data we use during our analysis is provided by the Bavarian Environmental Agency's hydrological service (GKD) (**?**) which is data from multiple measurement station along the Isar and the Danube. Based on this, the following gives a brief description of the data, discusses possible flood event detection methods, derives the variables of interest based on the flood definition and ends with a display of the crucial aspects of the obtained data.

Initially, the data contains discharge values in 15 minute steps for 27 stations along the Isar and Danube from different starting time points, but always up to 31.12.2024. We removed removed 6 measurement stations because these contained only a few observed years which is problematic because the final copula model is fit on yearly data. Thus, the number of observed years corresponds to the number of data points our copula model relies on.
Of the remaining 21 stations, 12 stations are along the Isar and 9 along the Danube where every station had at least 44 years of observation. As seen towards the end of this section, the alpine river Isar and the low-lying Danube have contrasting hydrological characteristics, enabling a meaningful comparison of flood dynamics in Bavaria. The exact spatial distribution of the considered station displayed plot 5.1.

Given the annual discharge data for all these stations, we require to identify the most severe flood event within each year which defined as the event with the largest discharge peak. To stabilize event detection, the

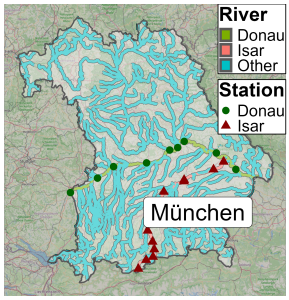


FIGURE 5.1: Caption

following is based on daily average discharge values we calculated based on the 15 minute time intervals in the original data.
The flood detection approach proposed by **?** of using the straight-line method based on a fixed threshold was found to be highly unreliable, but so was a quantile-based straight-line method. Both approaches exhibit significant uncertainties in identifying flood events, particularly, they tend to overestimate flood duration. Instead, we applied the baseflow methods proposed and implemented by **?**. This method relies on the baseflow index (BFI) which is the ratio of the baseflow volume to the volume of streamflow.

TODO: Definition baseflow and streamflow

A default BFI threshold of 0.5 was used to distinguish events dominated by rapid runoff contributions typically associated with rainfall- or melt-induced flooding.

Exemplary, figure 5.2 shows the hydrograph for the station in Munich in 2024.

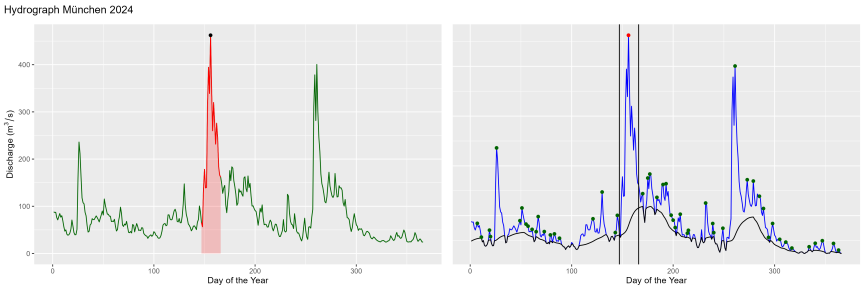


FIGURE 5.2: Caption

TODO: Depending on plot, describe what the F is seen here

Based on all identified flood events, the event with the largest peak discharge is selected and, finally, the variables of interst are determined. That is, flood peak is the maximum discharge value occurring within the event, flood duration is the time span measured in days between the start and end of the event, as determined by the BFI threshold crossings. Flood volume is the cumulative streamflow over the flood duration, representing total discharge volume in m^3.

TODO: Do I want to give some max values?

Now, we come to the most crucial aspect in our data, but show that this structure is also found in **?**. That is, fig 5.3 displays the rank correlation coefficient Kendall's τ between every possible combination between the 3 variables separated by river. In section 5.3.9 we further discuss Kendall's τ, but for now, it is sufficient to consider it as measure of the strength of the dependence between two variables.

The boxplots in figure 5.3 are based on the 9 and 12 stations along each river, respectively, and depict the τ values for the corresponding variable combination seen on the x-Axis. The black dots refer to the τ values observed by **?**. Most important here is that none of the boxplots align horizontally. That is, the strength of

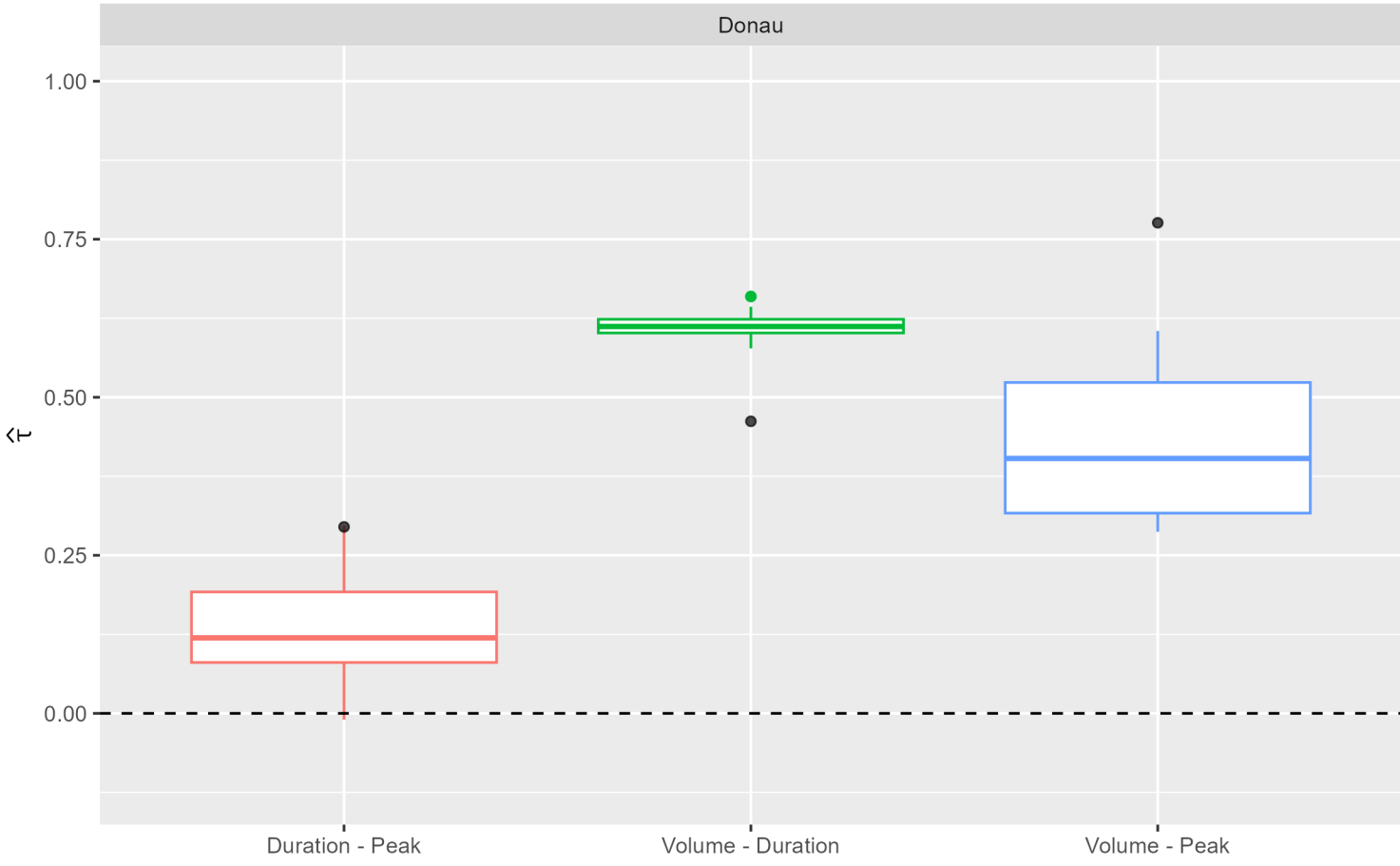


FIGURE 5.3: Caption

dependence differs between all pairs of variables. Thereby, our data suggests 3 different distinct dependencies. This finding is most crucial and, as we will see later on, renders **?** approach infeasible. Because, as seen from the black dots, not only our data suggest 3 separate dependence structures, but also the river **?** considered. Also interesting is the exact order of correlation values by each river. For both rivers, duration and peak always had the lowest correlation value. For the Danube, volume and duration are always the variables with the highest correlation with an exception of only one station. Nevertheless, all these values are quite similar as seen from the width of the boxplot in figure **??**. For the Isar, on the other hand, we observe not only more variation in the correlation values, but here the most correlated pair tends to be volume and peak. Of the 12 stations, 8 had volume and peak to be the pair with the highest correlation.
This emphazises the aforementioned contrasting hydrological characteristics which are highly relevant for copula modelling and, thereby, for our analysis.

Finally, the analysis section utilizes return periods of flood peaks to derive average discharge values conditioned on a certain peak. A return period is the average duration it takes for a peak value to re-occure and is usually measured in years. This period follows from the inverse probability of an event to take place. Thus, the peak value increases with an increasing return period. An average discharge value refers to the average discharge during a flood event. Thus, it is obtained by $\frac{volume[m^3]}{duration[s]}$ and to ensure these are comparable among stations, they are normalized by the station specific mean and standard deviation. To now characterize a conditional distribution using our data, we ordered all flood events within each station by their peak values and then selected the quantiles corresponding to the return periods $2, 5, 10, 20$ and 50 years. Thereby, we obtained 21 average discharge values for each return period. Consider figure **??** for a visualization.

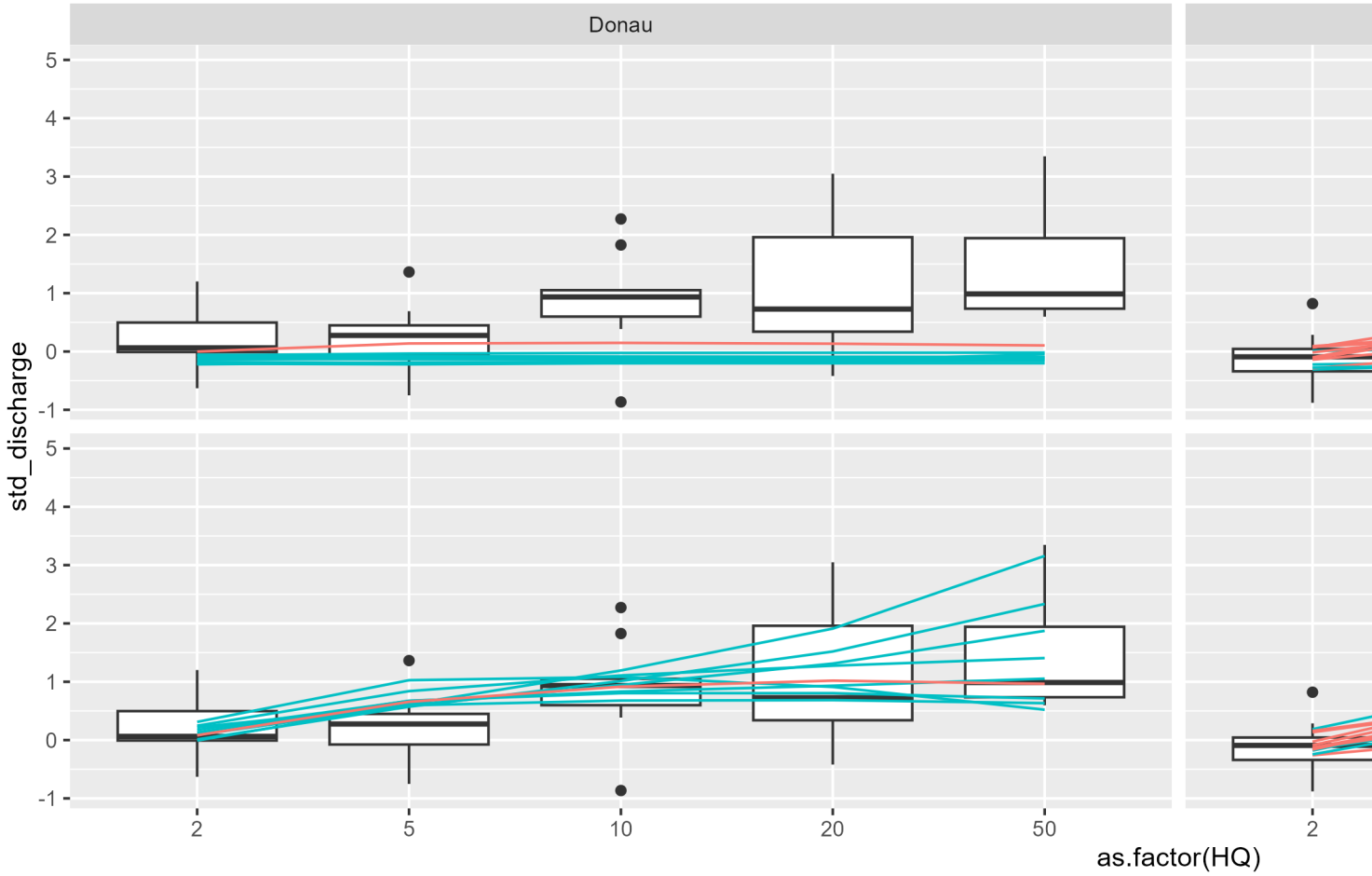


FIGURE 5.4: Caption

First of all, note that the average discharge values are on the y-axis ranging from −1 to 5. This is due to the standardization process within each station. The x-axis denotes the return periods of the peak based on the quantile approach. To account for the different structure between rivers, the figure considers them separately. Thereby, each boxplot in the Danube column is based on 9 data points and 12 for the Isar where each data point corresponds to a station. For both subplots, the average discharge increases with an increase in the return period. However, while the subplot of the Danube suggests a moderate increase in average discharge values, the Isar has a stark increase. But by the size of the boxplot at return period 50, this increase does not hold for all stations.

5.3 Methods

To address the dependence structures identified in the previous section, this chapter extends the approach of **?** by incorporating vine copulas. This extension is necessary because a simpler approach is insufficient to capture the full correlation pattern observed in the data. The following introduces the foundational theory of copulas, the family of Archimedean copulas as well as nested Archimedean and vine copula models. In addition, methods for copula fitting and model selection are briefly discussed. Then, some applied, but non-

essential methods are briefly established. Together, these elements form the theoretical framework on which this paper is based. Finally, a few words to the implementation of these methods and used packages.

5.3.1 Copulas

? (p. 62) describe a copula as a cumulative distribution function (CDF) with standard uniform margins. The dimension d of a copula denotes the number of random variables it relates and, hence, a copula is at least bivariate ($d \geq 2$). To give a mathematical definition, consider the vector $u = (u_1, ..., u_d) \in \mathbb{R}^d$ where $u_j \in [0, 1]$ for $j = 1, .., d$. Then, a d dimensional copula is defined by **?** (p. 14) as function $C : [0, 1]^d \to [0, 1]$ if, and only if, the following conditions hold:

i) $C(u_1, ..., u_d) = 0$ if $u_j = 0$ for at least one $j \in \{1, ..., d\}$.

ii) $C(1, 1, ..., 1, u_j, 1, ..., 1) = u_j$

iii) C is d-increasing

According to **?** (p. 9), condition i. shows that copulas are grounded. In this context, grounded means that plugging in 0 for just one of the variables yields a copula value of 0, independent of the other variables' value. The author also mentions that, using condition ii., the margins of the function C with respect to a certain variables are obtained by plugging in 1 for all other variables. Finally, the condition of C to be d-increasing is cumbersome to map out in higher dimensions, which is why the following is restricted to the $d = 2$ case. According to **?** (p. 8), the copula function C is 2-increasing if for all $u_1, u_2, v_1, v_2 \in [0, 1]$ with $u_1 \leq u_2$ and $v_1 \leq v_2$:

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$$

Simply put, 2-increasing means that the volume under the copula density function over the rectangle $[u_1, u_2] \times [v_1, v_2]$ is non-negative. This interpretation follows from the fact that copula functions are defined as CDF and holds for higher dimensions, too.
The next section introduces the central theorem in copula theory and also derives the already mentioned copula density.

5.3.2 Sklar's Theorem

Sklar's Theorem is central to the theory of copulas as it proves that any multivariate distribution can be constructed using copulas (**?** p. 17, **?** p. 42). Thereby, this theorem allows to separate the representation of the dependence structure and marginal distribution functions. The theorem is given by **?** (p. 18):
Let $F_{1,..,d}$ be a d-dimensional joint distribution function with univariate margins $F_1, ..., F_d$. Then, there exists a d-dimensional copula C such that

$$F_{1,...,d}(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d)) = C(u_1, ..., u_d)$$

where $u_i = F_i(x_i)$. Also, C is unique if $F_1, ..., F_d$ are continuous. Equation (5.3.2) allows 2 important conclusion: One, any multivariate CDF may be expressed as a composition of a copula function C and the univariate margins $F_1, ..., F_d$. Thereby, **?** (p. 66) conclude that C connects the multivariate CDF to its margins which allows to separately consider marginal and joint behavior of variables. That is, the problem of determining any multivariate CDF is reduced to determining the copula. And two, the marginal distributions do not need to be of the same family because Sklar's theorem holds regardless.

The aforementioned copula density function is given by (see **?**, p. 66):

$$c(u_1, ..., u_d) = \frac{\partial C(u_1, ..., u_d)}{\partial u_1 ... \partial u_d} = \frac{f(x_1, ..., x_d)}{\Pi_{i=1}^d f_i(x_i)}$$

where $f(x_1, ..., x_d)$ denotes the joint density of $X_1, ..., X_d$ and $f_i(x_i)$ the marginal density of X_i for $i = 1, ..., d$. Based on this equation, the joint density in terms of the copula density is given by

$$f(x_1, ..., x_d) = c(u_1, ..., u_d)\Pi_{i=1}^d f_i(x_i)$$

5.3.3 Symmetric Archimedean copulas and generator functions

As Nelsen (p. 109) states, symmetric Archimedean copulas (SACs) are widely applied due to their large variety and easy construction. However, SACs only allow the same dependence strength and structure among all possible pairs of variables as **?** (p.124) point out. Therefore, they are not suitable for our analysis as concluded from figure 5.3 which suggested 3 distinct correlation values. However, SACs remain an important building block for more complex copula models. Thus, this section introduces the concept of a generator function as it determines the family a SACs belongs to. Then, we specifically focus on bivariate SACs because following models are based on these.

We first give the general idea of a generator, then the representation of a copula in terms of the generator and in the end the copula families we use for our analysis.
? (p. 110, 111) defines a generator to be a continuous and strictly decreasing function $\phi : [0,1] \to [0,\infty)$ such that $\phi(1) = 0$. If $\phi(0) \to \infty$, the generator is considered to be strict. The inverse $\phi^{-1} : [0,\infty) \to [0,1]$ of such generators is strictly decreasing on $[0, \phi(0)]$. We only apply strict generators as seen towards the end of this section.
For a generator to yield a valid d-dimensional copula, **?** and **?** (p. 124) mention that the inverse requires to be completely monotone which is given if it has derivatives of all orders with alternating sign

$$(-1)^k \frac{d^k \phi^{-1}(t|\theta)}{dt^k} \geq 0.$$

Now, we are in the position to formulate the general representation of a d-dimensional SAC in terms of its generator. The relation is given by **?** (p. 123) as

$$C(u_1, u_2, u_3 \mid \theta) = \phi^{-1}(\phi(u_1 \mid \theta) + \phi(u_2 \mid \theta) + \phi(u_3 \mid \theta) \mid \theta).$$

Equation (5.3.3) shows that SACs are uniquely defined by their generator function and a parameter vector θ which we introduce next. As mentioned by **?** (p. 110, 111, 114), the assumed functional form of the generator translates to a specific copula family. Or, vice versa, assuming a copula family implies assuming a specific generator function. The θ vector, on the other hand, influences the dependence strength within the assumed copula family as seen in **?** (p. 86). This parameter vector takes on an important role in fitting a copula to observed data. That is, for an assumed copula family, this parameter vector remains to be estimated from the data. The exact approach is further discussed in section 5.3.7. For now, note that we focus on the 3 generator functions with a one-dimensional θ vector. These are specified in table 5.1.
Finally, equation (5.3.3) shows that the arguments to the SAC are exchangeable (see **?** (p. 38)). Exchangeability is a form of symmetry and implies that the copula treats all its arguments the same. Thereby, this representation displays the aforementioned restriction of SACs being able to only depict one unique dependence structure.

TABLE 5.1: Generator functions of selected Archimedean copulas according to **?** (p. 130) and tail dependencies according to **?** (p. 132).

Copula Family	Parameter θ	Generator Function $\phi(t)$	Tail Dep.
Clayton	$\theta \in [-1,\infty) \setminus \{0\}$	$\phi(t\|\theta) = \frac{1}{\theta}(t^{-\theta} - 1)$	Lower
Gumbel-Hougaard	$\theta \in [1,\infty)$	$\phi(t\|\theta) = (-\ln t)^{\theta}$	Upper
Frank	$\theta \in (-\infty,\infty) \setminus \{0\}$	$\phi(t\|\theta) = -\ln\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right)$	None

5.3.4 Taildependence and Rotation

After explaining what it means for a copula to be of a certain family, the following introduces the family specific concept of tail dependence. Also, we briefly explain how copulas are manipulated to extend the dependence possible structure one family captures.

Tail dependence is differentiated into upper and lower tail dependence. Their formulas are given by **?** (p. 34 - 35) as and As seen from both equations, tail dependence is defined as conditional probability that both

variables are above or below a threshold quantile. Thereby, tail dependence measures how likely it is for both random variables to jointly exhibit extreme behavior. However, upper tail dependence refers to both variables attaining large values while lower tail dependence means both variables are jointly small. Note that both, upper and lower tail dependence, depend on the copula function C and parameter θ and, thus, on the copula family. The tail dependencies implied by the copula families we consider are also listed in table 5.1 as stated in **?** (p. 132).
Finally, tail dependence is a family specific property, however, a copula function may be rotated to change its native tail dependence behavior. Following **?**, this is done by modelling $u_i' = 1 - u_i$ instead of u_i itself. Every such transformation concludes in a 90 degree rotation of the copula function in the corresponding direction. This approach is based on the definition of the copula function as multivariate CDF. Because if instead of u_i the transformation u_i' is modelled, the probabilistic statement of the copula in the continuous case changes to $C(u_1', u_2|\theta) = \mathbb{P}(X_1 \geq x_1, X_2 \leq x_2)$ and $C(u_1, u_2'|\theta) = \mathbb{P}(X_1 \leq x_1, X_2 \geq x_2)$, respectively.

5.3.5 Fully Nested Archimedean copulas

Fully nested Archimedean copulas (FNACs) build upon SACs and partially alleviate their restrictions. Note that these are the models **?** made extensive use of.
As our analysis applies to the trivariate case, FNACs and vines in section 5.3.6 are introduced for this trivariate case only.

FNACs are built by nesting bivariate SACs
$$C(u_1, u_2, u_3 \mid \theta) = C_1\left(C_2(u_1, u_2 \mid \theta_2), u_3 \mid \theta_1\right),$$
where θ_1 and θ_2 are the parameters corresponding to copula function C_1 and C_2 and $\theta = (\theta_1, \theta_2)$ is a vector containing all parameters. Note that there are only 2 distinct parameters θ_i which is why FNACs, in the trivariate case, are only able to capture 2 distinct dependence structures. This allows 2 conclusions. First, partial exchangeability remains which means that within the bivariate nested copulas C_2, the two arguments u_1, u_2 are interchangeable (see **?** p. 375). So to a degree, symmetry prevails. And second, the nested variables u_1, u_2 have the same marginal relation with u_3. That is, $C(a, 1, u_3|\theta) = C(1, a, u_3|\theta) = C_1(a, u_3|\theta_1)$. In essence, the statements are equivalent but the important take away is that FNACs are not able to display 3 distinct dependence structures. Thereby, they are not suitable for analysis. As also **?** observed 3 distinct correlations, their results are questionable for the same reason.
Additionally to this restriction, **?** mention that FNACs require the sufficient nesting condition to be fulfilled for equation (5.3.5) to yield a valid copula. We limit our considerations to FNACs where all nested copulas are of the same family. This corresponds to what **?** used in their analysis. Then, the sufficient nesting condition is fulfilled if deeper nested variables have a stronger degree of dependence, i.e. $\theta_1 \leq \theta_2$. (see **?**).

Note that equation (5.3.5) may also be represented in terms of the generator function. Thereby, additional requirements regarding the composition of generator functions emerge, as mentioned by **?** (p. 174). However, discussing these requirements is beyond the purpose of this paper. The interested reader is refered to **?** (p. 174).

5.3.6 Vine Copulas

Vine copulas use the pair-copula construction (PCC) explained by **?** (p. 77 - 80) to characterize multivariate dependence structures. That is, PCC decomposes multivariate densities into products of (conditional) bivariate densities (see **?** p. 88).
There exist multiple vine copula classes, depending on the structure the PCC implies, but as we are concerned with the trivariate case, these constructions are equivalent.

We use **?** (p. 78, 90) and devine the trivariate copula density as
$$c_{123}(u_1, u_2, u_3|\theta) = c_{12}(u_1, u_2|\theta_{12}) \cdot c_{23}(u_2, u_3|\theta_{23}) \cdot c_{13|2}(u_1|u_2, u_3|u_2 \mid \theta_{13|2})$$
where $u_i|u_j = F_{i|j}(x_i|x_j)$ denotes the conditional probability. Note that this copula density is based on the simplifying assumption for vines (**?** p.90, **?**) which means that the conditional bivariate density $c_{13|2}$ is

independent of exact x_2 values. It only depends on the conditional probabilities.
Visible from the number of parameters in the θ-vector, vines are able to capture all 3 distinct dependence structures in the trivariate case. Also, in contrast to FNACs from section 5.3.5 where we followed **?**, we do not require the bivariate copulas to be of the same family. Thereby, not only the strength of dependence between all 3 variables may differ, also the dependence structure is allowed to change from pair to pair.

5.3.7 Estimation and Selection Process

In practice, we need not only to estimate the parameter vector θ, but also select the best fitting copula. Thus, the following briefly introduces the pseudo maximum likelihood (ML) approach. Also, we give a reminder on the Akaike Information Criterion (AIC) as it is our information criterion of choice to select a copula model.

To avoid assumptions on the marginal distributions, we estimate the parameter vector θ using the pseudo-likelihood proposed by **?**

$$\hat{\theta} = argmax_\theta \; l(\theta) = argmax_\theta \sum_{k=1}^{n} log[c(u_{1k}, u_{2k}, u_{3k}|\theta)],$$

where u_{ik} denotes the marginal empirical distribution function scaled by $\frac{n}{n+1}$. These transformed variables are refered to as pseudo-observations. Depending on the copula model, the definition of the copula density follows either from equation (5.3.5) or (5.3.6). Thus, the exact estimation process slightly differs between copula models.
The AIC is given by **?** (p. 164) as

$$AIC = -2l(\hat{\theta}) + 2(|M| + 1)$$

where $l(\hat{\theta})$ represents the log-likelihood of the copula model fit and $|M|$ the number of parameters included in the model. We select that copula model with the smallest AIC.

5.3.8 Identifying univariate margins

While the the estimation process described in section 5.3.7 utilizes the empirical distribution function, an empirical function has undesired properties when re-transforming copula data. That is, during the estimation process, the empirical distribution functions ensures we do not affect the copula model fit by missspecifying the marginal distributions. However, after fitting the copula, whenever the inverse of the empirical distribution is applied, it bins any continuous data because it is a step-wise function. This, of course, limits the power of our copula analysis. Thus, we decided to fit a Generalized Extreme Values (GEV) distribution to the marginal distribution only for re-transforming any results from the fitted copula models.

The distribution function for the GEV family is given by **?** (p. 47)

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\},$$

defined for $\left\{z\middle|1 + \xi\left(\frac{z-\mu}{\sigma}\right) > 0\right\}$ where $-\infty < \mu < \infty$ denotes the location parameter, $\sigma > 0$ the scale parameter and $-\infty < \xi < \infty$ the shape parameter. In practice, these parameters are usually unknown, but **?** (p. 50) describes how these parameters are estimated from data using the ML approach.
Since GEV distributions are not at essence for our work, we refer the interested reader to **?** (chapter 3) for a more detailled consideration.

5.3.9 Kendall's τ

According to **?** (p. 6), τ is a measure of association between two random variables that distinguishes concordant and discordant pairs. Concordance means that the two variables move in the same direction while discordance means moving in opposite directions.
? (p. 86) and **?** (p. 159, 161 - 164) show that Kendall's τ is directly connected to the generator function and,

thus, a function in the parameter θ.

$$\tau(t) = 4 \int_0^1 \frac{\phi(t|\theta)}{\phi'(t|\theta)} dt + 1$$

where $\phi'(t|\theta)$ denotes the derivative of the generator function. Note that this relation is positive which is seen in **?** (p. 134). That is, if the parameter of copula increases, the strength of dependence increases. Vice versa, if correlation increases, θ increases, too. Also, note that this implies that estimating a θ implicitly estimates a τ value. This is a relation we will use during our simulation.

Empirically, there are multiple versions of Kendall's τ depending on the data structure. Since this paper focuses on continuous variables only, the formula given by **?** (p. 5) is applicable

$$t = \frac{P - Q}{\frac{1}{2}n(n-1)}.$$

P denotes the number of concordant and Q the number of discordant pairs in the data.

5.3.10 Software

The whole analysis is implemented using the programming language R. We used the `hydroEvents` package by **?** `eventBaseflow`. This method relies on the BFI as explained in the data section using a deftault BFI of 0.5. , Additionally, `eventBaseflow` function calculates the BFI at each time step and extracts discrete flood events when the BFI falls below the specified threshold for a user-specified minimum duration. For copulas, we relied on the `copula` package by **?** and the `pobs` function to apply the empirical distribution function as described in section 5.3.7. FNACs are dealt with using the `HAC` by **?**, especially the function. `estimate.copula` for FNAC fitting which implements the ML approach from section 5.3.7. This required some additional code to select the best fitting copula according to the AIC. For vine copulas, `VineCopula` package by **?** was consulted. Especially the function `RVineCopSelect` which implicitly fits a selection of copulas and also selects the one with the smallest AIC. Finally, GEV distribution were fitted using the function `fevd` from the `extRemes` package by **?** which uses MLE to determine the parameters mentioned in 5.3.8.

5.4 Simulation

To examine the incapability of FNACs to capture 3 distinct dependence structures in a trivariate setting, we ran a simulation. The true underlying model builds a vine copula models. This section is limited to the most crucial finding which is highly relevant for the interpretation of our results in the following section.

We set up the simulation by drawing 27000 random samples of size $15, 30, 50, 1000$. The sample size of 50 represents our real world conditions while 1000 observations aim to examine large sample behavior. The two smaller samples sizes are interesting because we earlier decided to remove stations due to their small sample size.
For each drawn sample, the underlying vine copula model has 2 wheels to tweak: First, each copula density of the underlying vine copula model is allowed to be of one of the copula families listed in table 5.1 leading to a possible total of $3^3 = 27$ copula family combinations. Second, we allowed for 4 different correlation values, 2 of which correspond to the observed average correlation values in the Isar and the Danube. The remaining 2 aimed to examine general behavior of FNACs. That is, we added a Low-Medium-High correlation structure with correlation values of $0.1, 0.5, 0.85$ and a Low-High-High structure using $0.1, 0.8, 0.8$.
Thereby, we had 27000 data points per sample size which are split among all 27 possible copula family combinations and 4 possible correlation structures. This leads to roughly $\frac{27000}{4 \cdot 27} = 250$ data points per setup. It is not exactly 250 because we used a uniform draw to select copula families and correlation structure as it drastically simplified implementation.

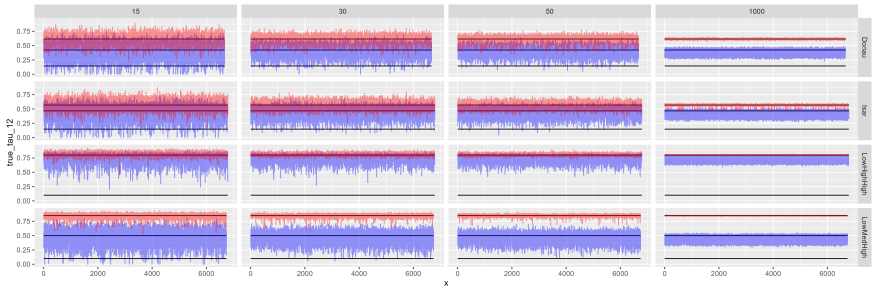


FIGURE 5.5: Caption

The one result we want to focus on is described in figure **??**.

This figure uses multiple subplots displaying trace plots of the τ values estimated by the FNAC model. Kendall's τ estimates are displayed on the y-axis while the index of the iteration in which this model was fitted is on x-axis. Note that this plot does not differ by copula family combination leading to roughly $\frac{27000}{4} = 6750$ iterations each. The black lines in each subplot refers to the true underlying correlation values. The name of the corresponding correlation structure is to the right hand side of the plot. Additionally, the figure is divided by the sample size on which the estimated FNAC model is based. From a sample size of 15 on the left up to a size of 1000 on the right. Color-wise, the red line refers to the τ of the nested FNAC and the blue line to the outer estimate. Due to the sufficient nesting condition (see section 5.3.5), the inner τ in each iteration
is always larger than the outer.
The first observation is the decrease in variance of the estimators as the sample size increases. This holds for all possible correlation structures and was expected. Now, focus only on the column of sample size 1000. Especially in the Low-Medium-High subplot, we observe that the inner τ estimate moves around the most upper black line. Thereby, the nested copula in a FNAC correctly captures the largest correlation value. The blue line, however, moves around the second largest black line and its volatility remains comparably large. We conclude that, first, the τ estimate of the outer copula of a FNAC model varies within a set interval for large sample sizes. This is similar to what **?** found in their work when they examined how SACs perform if the true underlying model is a FNAC. Second, and highly relevant for our analysis, the τ based on the outer copula of a FNAC tends towards the second largest correlation value in the (simulated) data. This implies that, due to the same copula margins mentioned in section 5.3.5, FNACs systematically overestimate the weakest dependence strength. This is an important result because it explains not only the comparably bad performance of FNACs during our application, but also their bias in the results.

5.5 Application

Due to our finding during the simulation, we focus on presenting the fitted vine copulas models. Only for the predictive performance, we jointly consider FNACs and vines to discuss the effect of the bias in FNACs and its practical meaning.

5.5.1 Visual Goodness of Fit

To each of the 21 stations, we fitted every possible copula family combination also allowing for a copula function to be rotated. Then, the best fitting copula, according to the AIC, is selected. Thereby, the following discusses the one best vine copula fit we found. Exemplary, we visually assess the goodness of fit for the Munich station using figure 5.6.

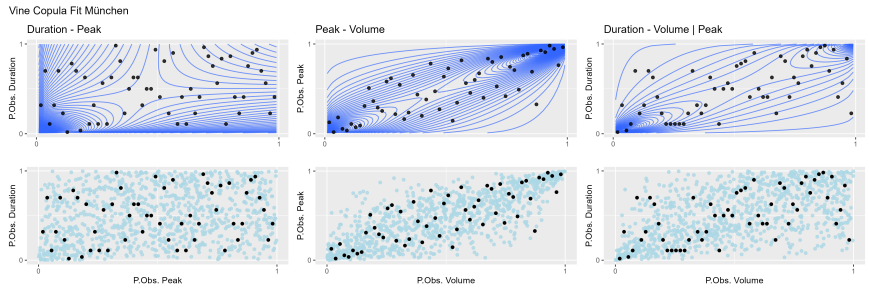


FIGURE 5.6: Caption

This figure consists of 6 subplots. Each column of subplots refers to a pair of variables specified in the column header. The variable named first is displayed on the y-axis. Note that all subplots are on copula level. Thus, all axes range from 0 to 1. The top row of subplots shows the contour lines for the fitted copula density and the bottom row a synthetic random sample from the copula model. The black points in every subplot depict the pseudo observations of the corresponding variables.
Starting with bottom row, a good model fit implies that the light blue points capture the structure of the pseudo observations. This seems to be the case for all variable pairs as not only the shape of the black points is nicely reflected in the synthetic data points, but also the strength of dependence matches. This is validated when comparing the empirical correlation values with the correlation implied by the fitted models as seen in table **??**. The largest absolute difference between the correlation values is just 0.02.

TABLE 5.2: Describtives on the fitted copula model for Munich station.

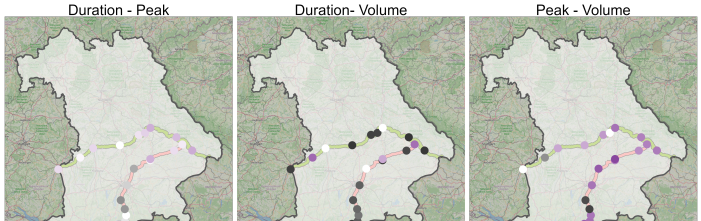
Pair	Copula Family	Empirical τ	Fitted τ	$\lvert\Delta\tau\rvert$
Duration – Peak	Clayton	0.15	0.16	0.01
Peak – Volume	Gumbel-Hougaard	0.49	0.51	0.02
Duration – Volume	180° Rotated Gumbel-Hougaard	0.60	0.59	0.01

Now, consider the top row of figure 5.6 in combination with the corresponding copula families mentioned in table **??**. The Clayton and 180° rotated Gumbel-Hougaard copula imply lower tail dependence for the relation between duration and peak as well as duration and volume, respectively. Thereby, if the flood peak is small, the duration tends to be short, too. Also, the volume of a flood is more likely to be small given a short flood duration. However, due to its small strength of dependence, the tail dependence for duration and peak is rather small. This is not only suggested by the contour lines, but follows from the low τ value which implies that paramter θ is small, too, as mentioned in section 5.3.9. This in turn affects the tail dependence as it is a function in θ, as discussed in section 5.3.4. Finally, a Gumbel copula is fitted to the peak and volume pair implying upper tail dependence. Thus, given a large peak, the flood volume tends to be large, too.

5.5.2 Fitted Tail Dependencies

Because tail dependence is an important concept from a hydrological point of view, we extend the tail dependence analysis from the previous section to all considered stations.

Contemplate figure 5.7 for a visual assessment of the tail dependence structure.



dependence values remain unchanged, but lower tail dependencies are multiplies by -1. Thereby, the blacker a station is colored in, the stronger the lower tail dependence and the more purple the stronger the upper tail dependence. A totally white point on the map refers to no tail dependence. In these cases, a Frank copula has been fitted to the variable pair.
In general, we observe lighter colors for the duration and peak pairs which is due to the rather small correlation values. This follows from the reasoning in the previous section. Duration and volume exhibit the darkest colors suggesting that if the flood duration is short, the flood volume tends to be small, too, for the majority of the considered stations.

Especially towards the alpine regions, the plot suggests a spatial trend.

TODO: Any hydrologically intersting reasoning for this? Slope? Mention that our slope data was shit or smth?

Finally, especially for the Isar, peak and volume seem to exhibit upper tail dependence. Thereby, a high peak is more likely to appear with a high flood volume for the stations along the Isar.

TODO: Any hydrological reason here?

5.5.3 Event Probability

The following examines the difference between a univariate and the multivariate approach for flood characterization and thereby addresses one of our research questions. To answer it, the following focuses on the flood event in Munich 2024 which had a peak of $462m^3/s$, took 20 days and had a volume of $406m^3$.

Figure **??** displays the marginal fit of a GEV distribution to the peak data for the Munich station.

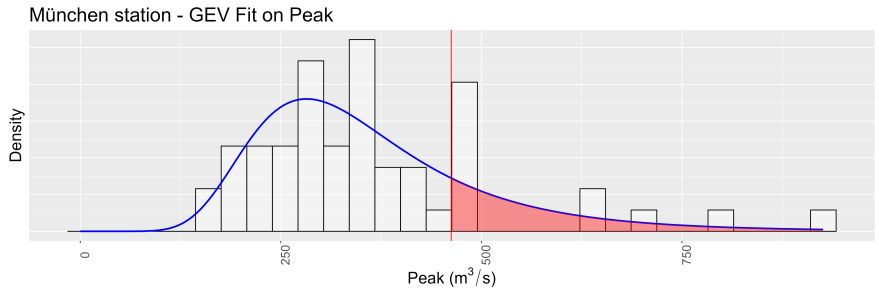


FIGURE 5.8: Caption

The x-axis of this plot denotes the peak values, the y-axis the density of the fitted GEV model. The histogram displays the original data to which the GEV distribution was fitted. The resulting smooth distribution is marked by the blue line. The red vertical line marks the peak value of $462m^3/s$ and the the shaded area visualizes the probability to observe a peak at least as large. Based on the GEV fit, the probability is calculated to be 19%. Thereby, this approach assigns the a return period of such a peak to be $\frac{1}{0.19} \approx 5$ years. Thus, Characterizing the whole flood event only by its peak, the return period of the whole flood in Munich of 2024 according to the univariate model is 5 years.

The trivariate copula model allows to also consider volume and duration values to characterize a flood event. Figure 5.9 is based on figure **??** and visualizes how the multivariate model determines the probability for a flood event to be at least as severe. Also, this plot helps to understand the quite stark differences in the probabilities.

Note that, as before, this figure displays the data in terms of their pseudo observations. The red dot in each subplot marks the event in 2024 and the red lines correspond to the univariate univariate observed pseudo values. Thereby, a flood event is at least as severe as the flood in Munich if it lies wihtin the shaded area. Thus, the probability is obtained by integrating the joint copula density over the cube made up by the shaded area. According to our model, this adds up to be 2.7% which corresponds to a return period of

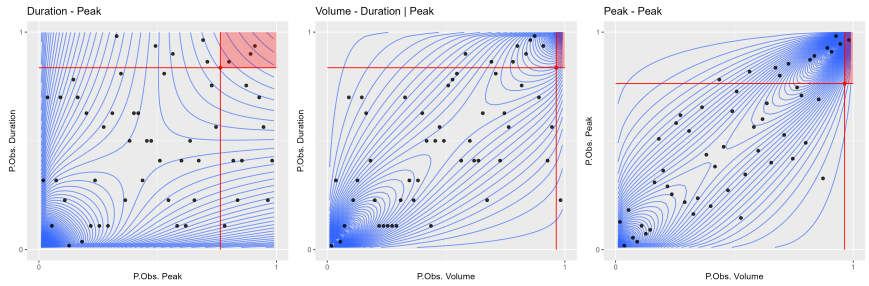


FIGURE 5.9: Caption

merely $\frac{1}{0.027} \approx 37$ years. Thereby, the return period for such an extreme flood is $\frac{37}{5} \approx 7$ times longer if the flood is characterized not only but its peak, but also by its volume and duration value. The reason for this drastic difference is seen in figure 5.9. While the pseudo observation of the peak value of the event is around 0.76, the volume is at 0.96. That is, the volume during this flood was exceptionally large which decreases the probability of such an event to occure. Visually, this is seen from the shaded area being very slim. In contrast, the univariate consideration of peak values only is not capable to account for this.

5.5.4 Conditional Discharges

This any good?:

Finally, this section examines the effect of different peak values onto the other characteristics of a flood. Mainly, we are interested in the average discharge value during a flood event as it is a measure of the average energy the system has to deal with.

First, we fit station specific GEV distributions onto the observed peak values and determine station specific peak values that only occure every $2, 5, 10, 20$ and 50 years. We choose the upper bound of 50 years to validate the model predictions using the available data. Conditional on these peaks, the most likely combination of duration and volume are determined. From these, the average discharge values are calculated and standardized in the same manner explained in section 5.2. Figure 5.10 is based on figure **??** and displays our results.

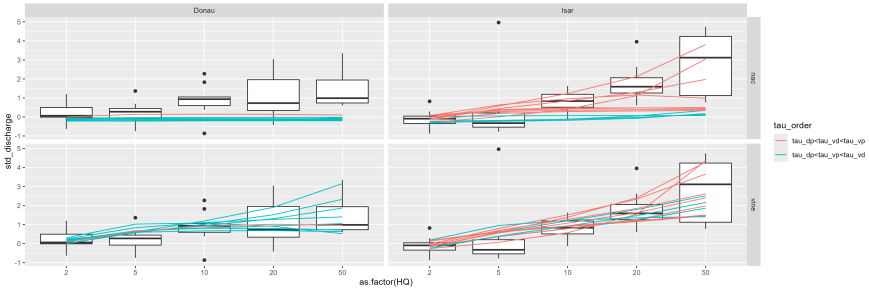


FIGURE 5.10: Caption

The rows of the figure refer to the model structure applied to predict the average discharges. Here, we also consider FNAC models because this application nicely depicts their shortcomings and reason why they are not suitable if their assumptions are violated. Additionally, we colored each model prediction by which variable pair had the highest correlation. First, consider the vine model fits in the bottom row. The models correctly capture the trend suggested by the boxplots independent of which pair of variables has the highest correlation. For a return period of 50, model predictions and boxplots increase in variance.

Does this make sense, hydrologically?: This is reasonable because the manner in which an event is extreme depends on the station. Thereby, a joint behavior in boxplot and model prediction suggests a good fit.

For FNACs, consider their performance within the stations of the Isar first. Their is a visible difference in model performance depending on which variable pair has a larger correlation. That is, if volume and peak build the inner copula due to their larger correlation, the model performs better than for volume and duration being stronger correlated. For the Danube, FNAC models bare move at all failing to capture any trend in the data whatsoever.

TODO: Where do we explain why this is the case? (FNACS failing) TODO: Systematic underestimation of average discharge bc bias

5.6 Discussion

From Probabilty section: As this section shows, we urge to not solely rely on a marginal GEV distribution of peak to fully characterize a flood event. Instead, a multivariate event consideration is required.

Discuss flood detection and effect onto results?

Use the fitted models to predict the most likeli discharge for some HQ100

6

Introduction

Author:

Supervisor:

6.1 Intro About the Seminar Topic

6.2 Outline of the Booklet

7

Introduction

Author:

Supervisor:

7.1 Intro About the Seminar Topic

7.2 Outline of the Booklet

8

Introduction

Author:

Supervisor:

8.1 Intro About the Seminar Topic

8.2 Outline of the Booklet

9

Acknowledgements

The most important contributions are from the students themselves. The success of such projects highly depends on the students. And this book is a success, so thanks a lot to all the authors! The other important role is the supervisor. Thanks to all the supervisors who participated! Special thanks to Helmut Küchenhoff[1] who enabled us to conduct the seminar in such an experimental way, supported us and gave valuable feedback for the seminar structure. Thanks a lot as well to the entire Department of Statistics[2] and the LMU Munich[3] for the infrastructure.

The authors of this work take full responsibilities for its content.

[1] https://www.stablab.stat.uni-muenchen.de/personen/leitung/kuechenhoff1/index.html
[2] https://www.statistik.uni-muenchen.de/
[3] http://www.en.uni-muenchen.de/index.html

Bibliography

Coles, S. (2001). An introduction to statistical modeling of extreme values.

Czado, C. (2019). Analyzing dependent data with vine copulas. 222.

Durante, F. and Sempi, C. (2016). *Principles of copula theory*. Chapman and Hall/CRC.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). Regression: Models, methods and applications. *Regression: Models, Methods and Applications*, 9783642343339:1–698.

Genest, C., Ghoudi, K., and p. Rivest, L. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82:543–552.

Gilleland, E. and Katz, R. W. (2016). extremes 2.0: An extreme value analysis package in r. *Journal of Statistical Software*, 72:1–39.

Grimaldi, S. and Serinaldi, F. (2006). Asymmetric copula in multivariate flood frequency analysis. *Advances in Water Resources*, 29:1155–1167.

Górecki, J., Hofert, M., and Holeňa, M. (2016). On structure, family and parameter estimation of hierarchical archimedean copulas. *Journal of Statistical Computation and Simulation*, 87:3261–3324.

Hofert, M., Kojadinovic, I., Maechler, M., and Johanna G. Nešlehová, R. M. L. Y. (2025). Package 'copula'.

Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., and Erhardt, T. (2024). Vinecopula: Statistical inference of vine copulas. R package version 2.6.0.

Nelsen, R. B. (2006). An introduction to copulas. *An Introduction to Copulas*.

Okhrin, O. and Ristig, A. (2014). Hierarchical archimedean copulae: The hac package. *Journal of Statistical Software*, 58:1–20.

Pan, R., Nieto-Barajas, L. E., and Craiu, R. Bivariate temporal dependence via mixtures of rotated copulas.

Vatter, T. and Nagler, T. (2018). Generalized additive models for pair-copula constructions. *Journal of Computational and Graphical Statistics*, 27:715–727.

Wasko, C. and Guo, D. (2025). hydroevents: Extract event statistics in hydrologic time series. R package version 0.12.0.

Zhang, L. and Singh, V. P. (2019). Copulas and their applications in water resources engineering. *Copulas and their Applications in Water Resources Engineering*, pages 1–603.