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

[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 **??**.

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

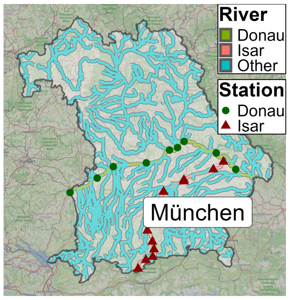

FIGURE 5.1: Caption

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 **??** shows the hydrograph for the station in Munich in 2024.

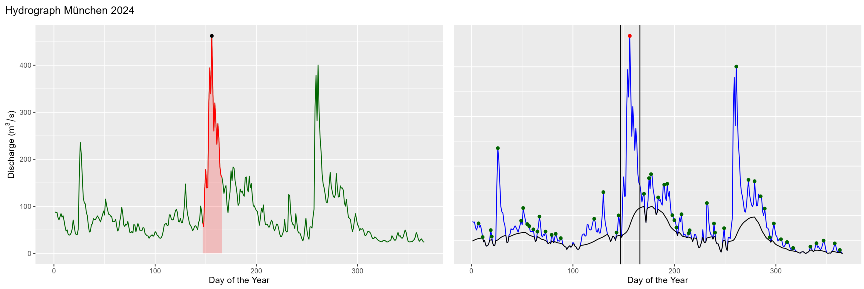

FIGURE 5.2: Caption

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?

Finally, we come to the most crucial aspect in our data, but show that this structure is also found in **?**. That is, fig **??** displays the rank correlation coefficient Kendall's τ between every possible combination between the 3 variables separated by river. In section **??** 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 **??** 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 dependence

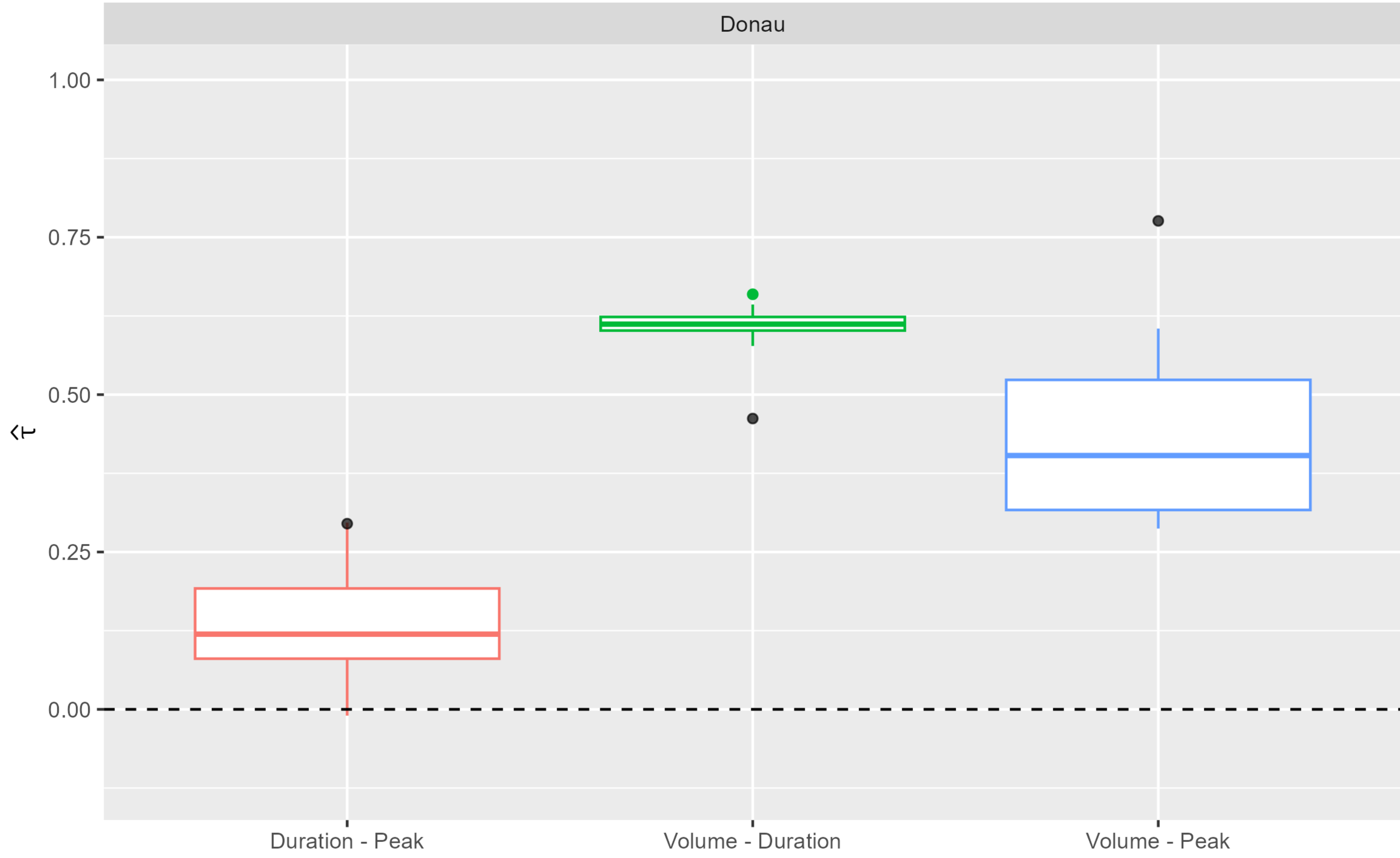

FIGURE 5.3: Caption

strength between all variable pairs differ to the others. 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 difference between the rivers in what variables have the strongest dependence. According to the figure, volume and duration are highly correlated with only little variation seen from the width of the boxplot. 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. This emphazises the aforementioned contrasting hydrological characteristics which are highly relevant for copula modelling and, thereby, for our analysis.

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 $C : [0, 1]^d \rightarrow [0, 1]$ 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 (**??**) 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

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