GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

A COMPARISON OF SPACE-TIME MODELING APPROACHES FOR EXTREME PRECIPITATION IN GERMANY BETWEEN 1996 AND 2016

A thesis paper for the degree of Bachelor of Arts (B.A.)

Wintersemester 2022

By

Ekin Gülhan

Supervisor: Dr. Isa Marques
Enrolment number: 21675680
e-mail: ekin.guelhan(at)stud.uni-goettingen.de
Address: Hannoversche Straße 134, 37077 Göttingen
Date: 2023-03-03

Contents

1 Abstract

2 Introduction

2.1 The need for a scientific approach to flood risk management and the purpose of the thesis paper

In 2013, massive floods resulting from continuous extreme precipitation hit Central Europe, causing loss of life and property in Germany and other affected areas. During this crisis, German institutions responsible for flood risk management took action and politicians publicly announced and evaluated the state of the crisis and countermeasures. However, critical voices also arose, asserting that the responsibility for the losses partially lay with the political failure of the institutions and officials at that time. Research conducted in the aftermath of the floods of 2013 evaluated the preparation for and response to the disaster, indicating the negative effect of political structures on the performance of flood risk management, including the dissemination of flood-related information through government agencies, which municipalities moreover found difficult to interpret and apply. Therefore, it is important to scientifically address questions related to improving local access to crucial information about the time and place of extreme precipitation events (EPE) to prevent such disasters in the future.

2.2 The design of this thesis project: Science, Data Science and Reproducibility

Against this background, it is the aim of the present thesis to determine which of two spatio-temporal classifier models performs better in predicting EPE in Germany. Observations of precipitation height between 1996 and 2012 have been used to train and evaluate the classifier models, while observations between 2013 and 2016 are meant for model validation and comparison. As the author aspires to a career in data science, the focus has been on not only applying scientific but also data scientific best practices, making reproducibility of the analyses a major concern.
For this reason, the thesis has been produced within a rigorously encapsulated and mobile project directory that can be accessed via a **GitHub repository**[1]. The R package "here" has been used

[1] Access to the repository on request. Reach out to me via ekin.guelhan@posteo.de

to enable mobility by substituting absolute file paths with relative file paths. This way, the path is defined relative to the topmost folder where the R project file resides. The R project file is the center of the practical implementation of the thesis. Using relative paths, different parts of the data manipulation, modelling, and analysis have been isolated in scripts that can be executed independently of the markdown file from which this thesis is knitted.

As is usually the case with R, computational functionality in those at least partially scripts relies on R packages provided by members of the R community. Reproducing the analytical steps, however, requires not only that the packages are installed but also that they are installed in a specific version. To minimize the effort needed for such package version control, the R package "renv" has been utilized, which creates a project package library secondary to the system library. A "renv.lock"-file is included in the project repository, which is a reference to all the required packages (and their specific versions) for reproducing the thesis and its analyses. Using the renv package, anyone with access to the thesis project repository can easily install the required packages via the "restore()" function.

Since the author considers the tidyverse as the current best practice of data manipulation and programming of data streams in R, packages associated with it, and in particular tidyverse pipes, have been rigorously used. Finally, for readability, vectors are d enoted through bold font, while matrices are denoted through bold font plus a description of the dimension as a subscript within mathematical expressions used in this thesis.

3 Framing the problem of spatio-temporally predicting EPE as a binary classification problem

As a preliminary to modeling the spatio-temporal prediction of EPE, a conceptual grounding of EPE is necessary. However, since this is not a priority in this analysis, such grounding will only be attended to as necessary. Indeed, it is not outright obvious when a precipitation event (PE) is to be regarded as extreme. A plethora of approaches have their common origin as early as making a choice of what constitutes a PE. Generally, researchers define a PE based on precipitation height, duration, or intensity, or based on a specific combination of the three.

Here, a PE is defined by the daily precipitation height only, as it is believed to be a potent predictor of the environmental impact and damage to infrastructure in itself. Among the possible temporal resolutions, the choice fell on days and against sub-daily resolutions (hourly precipitation) or more aggregated resolutions (monthly or annual precipitation) for reasons persistent with the previous argument: Daily precipitation totals are likely better predictors of the environmental impact of precipitation than other resolutions, which ultimately is the primary concern for local risk managers. These two choices for essentially basing the concept of PEs on daily total precipitation height have already cut away much of the variety of approaching the conceptual grounding.

At this point, where a daily PE is characterized by the daily precipitation height associated, to also complete the definition of EPE, a threshold (or cut-off) value needs to be decided upon for a specific standard of truth, or an axiomatic unambiguous way to seperate PE from EPE, respectively. Hence the use of the threshold value is to mark a threshold or boundary, based on whom each PE with its associated precipitation height, be it observed or predicted, is declared as either EPE or non-EPE. This declaration as either EPE or non-EPE is then finally be framed as a classification, with one class consisting of all the EPE, and the other class consisting of non-EPE, such that all PE can unambiguously and exhaustively be classified as either of both. The value of this conceptual act of framing the task of predicting EPE as classifying a given PE as either EPE or non-EPE, which is in fact a simple matter of definition, lies with the literature concerning the modeling, building / training and evaluation of classifier models, which hence becomes applicable. Thus, for the concrete task of establishing a threshold value, a specific percentile of the empirical distribution of precipitation may be considered in the first instance. This approach is indeed being

applied in the literature in some cases. Obvious related shortcomings, however, are the lack of scientific research that indicates that any specific quantile of the empirical distribution performs well in classifications, that reliably predict risks of environmental hazard.

Alternatively, there is a rich whole area of statistics to draw concepts for the cut-off value from, and it is called extreme value theory. This theory has a more differentiated view of the analysis of the values at the precipitation height distribution's tail. Attempts to tie it into this thesis, however, would have increased the difficulty unnecessarily. Hence, the choice for the threshold value fell on an easily comprehensible yet powerful concept:

Let $P_{(lon_i,lat_r,y,m,d)}$ be the daily total precipitation height observed at spatial location (lon_i, lat_r), where lon_i and lat_r denote the longitudinal and latiduinal degree respectively and y, m and d denote a year, month and day of the month. Also, let $P_{lon_i,lat_r,y,m}$ be the total precipitation height of the month m in the year y.

Then, the monthly total precipitation height $P_{(lon_i,lat_r,y,m)}$ for location (lon_i, lat_r) in year y can be calculated as:

$$P_{(lon_i,lat_r,y,m)} = \sum_{d \in m} P_{(lon_i,lat_r,y,m,d)} \tag{1}$$

For the same spatial location (lon_i, lat_r), the average of all monthly total precipitation heights for the month m observed in the different years in the training data $y\ in\ Y_{training} = \{1996, 1997, \cdots, 2012\}$ can then be calculated as:

$$\overline{P}_{(lon_i,lat_r,m)} = \frac{1}{N_{lon_i,lat_r,m}} \sum_{y \in Y} P_{(lon_i,lat_r,y,m)} \tag{2}$$

where $N_{lon_i,lat_r,m}$ is the number of years y, for whom precipitation height has been observed in the month m at the location (lon_i, lat_r).

The equation 2 calculates the average monthly total precipitation height of month m at location (lon_i, lat_r) across all years y, which represents the typical total precipitation height for that month at that location.[2]

[2]Note that the monthly typical total precipitation height is very likely unique for each spatial location.

A PE then is classified as an EPE, if:

$$P_{(lon_i,lat_r,y,m,d)} > \overline{P}_{(lon_i,lat_r,m)} \tag{3}$$

The standard of truth for the classification of a PE as an EPE hence is:

$$EPE_{(lon_i,lat_r,y,m,d)} = \begin{cases} 1, & \text{if } P_{(lon_i,lat_r,y,m,d)} > \overline{P}_{(lon_i,lat_r,m)} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where EPE is a binary variable that indicates whether the PE is classified as an EPE or not.

4 A classifier based on a GAMM with a logit-link for the spatio-temporal classification of EPE

The following section will argue that, on a conceptual level, the problem of modeling the spatio-temporal classification of PE as EPE can be approached reasonably by building a generalized additive mixed model (GAMM) with a logit link function, which is then superimposed by a filtering probability threshold layer. The threshold layer superimposing the built GAMM is ultimately responsible for separating the EPE from the PE based on the criterion of whether the predicted probability exceeds the threshold layer or not.

In this context, "reasonably" refers to the fact that building a classifier based on a GAMM can meet all the essential criteria for the model, which follow from the general objective of this thesis to build a classifier model that facilitates flood risk management, as follows:

Firstly, the classifier model has to be capable of translating an input of the spatial and temporal location of the PE into a classification decision. This relates to the presence of both the spatial and temporal location of the event as covariates in the model. Modelling the classification spatio-temporally also necessitates identifying spatial, temporal, and spatio-temporal dependencies within the data, since neglecting those dependencies would result in structurally incorrect classification decisions/prediction errors.

Next, in addition to the spatio-temporal character of the model, it should also be possible to train the classifier to identify non-linear relations between the classification decision and the covariates. Research indicates that the effect of variables such as altitude of spatial locations and mean temperature, which are generally considered strong predictor variables, is non-linear.

Finally, a classifier model involves a binary response, where in this case, a predicted value of 1 translates into "PE in class of EPE," whereas a predicted value of 0 translates into "PE not in class of EPE." The respective model must, therefore, be able to handle a binary response variable conceptually.

One way to show how a classifier based on a GAMM with a logit link meets all these criteria is to start with the classical linear regression model. Although the classical linear regression model only meets one requirement of modeling the dependency of PEs on spatial and temporal covariates, it can be shown that following successive interventions within the model, step by step more of the

requirements for the classifier model are met, such that the classifier based on the GAMM with logit-link is ultimately developed conceptually as a model to meet all criteria.

4.1 Classical Linear Regression and its shortcomings

Classical linear regression involves modeling the model response as the conditional expected value of a variable, with this expected value being dependent on one or more covariates in a linear fashion, plus an error term.[3] The random error terms are assumed to be identically and independently Gaussian-distributed. The covariates form a linear predictor such that, based on the modeled dependence of the response on both the random errors and the linear predictor, the response is conceptualized as a random variable as well, whose distribution, beyond adopting the i.i.d. features, is additionally influenced by the linear predictor. Importantly, as the error terms conceptually account for random, unsystematic differences between the response and the linear predictor, as well as measurement errors, they are assumed to be independent not only of themselves but also of the linear predictor. These characteristics collectively render linear regression fundamentally inadequate for the present purpose. Firstly, the response is assumed to be Gaussian-distributed, whereas for a binary response, such as a classification decision derived from a probability threshold layer, a binomial distribution needs to be assumed. Secondly, strong evidence exists for the existence of spatial, temporal, and spatio-temporal dependencies between precipitation heights and within the conditional dependency of the precipitation height of the covariates on one side and the assumed independence among the responses and among the error terms on the other side. If the responses (the classification decisions) are non-independent (autocorrelation), then these dependencies need to be captured through the linear predictor or the error terms. As the linear predictor, however, is inadequate for modeling any other than linear effects of the covariates on the responses, these dependencies would conceptually be assumed to be reflected in the error terms. The dependencies would therefore reflect in the error terms, particularly as autocorrelation of the error terms, which contradicts the assumption of i.i.d. error terms. Hence, as linear regression is, by construction, inadequate to meet the requirements of modeling a binary response and modeling spatio-temporal dependencies, and as this inadequacy finally also extends to modeling non-linear dependencies,

[3] For the model equation, see equation 10

linear regression overall is ill-suited for modeling the classification of PE as EPE, as this thesis is set out for.

4.2 Logistic Regression Models as *Generalized* Linear Models

The classical linear regression model can still generate predictions from the spatial and temporal location of a PE. And with changes to the model structure, it can be transformed into a probability estimation model. Although this logistic regression model does not model the binary variable of classification decisions directly, it generates estimations for the probabilities, that a PE is classified as an EPE. To take the final step to model the binary classification variables finally as a model response that depends on a linear predictor then would only require superimposing a probability-threshold layer onto the predictions of the logistic regression model.

But to achieve such a generalization of the linear model towards the logistic regression model, firstly the error term is dropped from the right-hand side, so that compared to linear regression, the model response is no longer modeled as dependent on the linear predictor plus an error term, but only on the linear predictor.

The core acts of the generalization then, which effectively allow to model classification probabilities $p(PE_{(lon_i,lat_r,jul_a)}\, in\, class\, EPE) = p_{(lon_i,lat_r,jul_a)}$ as dependent on a linear predictor, involve

- observations of these probabilities in the data
- a logit-function $logit(p_{(lon_i,lat_r,jul_a)}) = \log\left(\frac{p_{(lon_i,lat_r,jul_a)}}{1-p_{(lon_i,lat_r,jul_a)}}\right)$, which projects the observed probabilities onto log-odds(logit) and
- a sigmoid function S of the form $S\left(logit(p_{(lon_i,lat_r,jul_a)})\right) = \frac{1}{1+e^{logit\left(p_{(lon_i,lat_r,jul_a)}\right)}} = p_{(lon_i,lat_r,jul_a)}$, that is the inverse of the logit-function and returns probabilities with log-odds as input

and the concept is as follows:

It need to be noted first, that the primary concern with modeling a binary response per linear regression had been, that it results in systemic error. The logit-function solves this particular problem by projecting the probability values on a continuous scale from $-\infty$ to ∞, such that

the resulting continuous log-odds can be modeled as dependent on a linear predictor. Since the log-odds are not the variable of interest however, the logit's inverse function, S, is applied on both sides of the model equation. As a consequence, within the logistic regression model, the variable of interest, the probabilities $p_{(lon_i,lat_r,jul_a)}$ are modeled as dependent on the inverse logit function S with the logg-odds as the model argument. And since the log-odds were modeled as dependent on a linear predictor in the first place, the model equation presents itself as

$$\begin{aligned} p(PE_{(lon_i,lat_r,jul_a)} \text{ in class EPE}) &= \mathrm{S}\left(\log\left(\frac{p_{(lon_i,lat_r,jul_a)}}{1-p_{(lon_i,lat_r,jul_a)}}\right)\right) \\ &= \mathrm{S}\left(\begin{array}{l} \beta_0 + \beta_1 \text{Mean Temperature}_{(lon_i,lat_r,jul_a)} + \beta_2 \text{Altitude}_{(lon_i,lat_r,jul_a)} \\ + \beta_3 \text{Longitude}_{(lon_i,lat_r,jul_a)} + \beta_4 \text{Latitude}_{(lon_i,lat_r,jul_a)} \\ + \beta_5 \text{Julian Date}_{(lon_i,lat_r,jul_a)} \end{array}\right) \end{aligned} \tag{5}$$

where

$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ are *estimated* per *Maximum Likelihood Estimation*

$$\begin{aligned} p_{(lon_i,lat_r,jul_a)} = \Bigg(1 + e^{-(\beta_0+\beta_1 MeanTemperature_{(lon_i,lat_r,jul_a)}+\beta_2 Altitude_{(lon_i,lat_r,jul_a)}+\beta_3 Longitude_{(lon_i,lat_r,jul_a)})} \\ \times e^{-(\beta_4 Latitude_{(lon_i,lat_r,jul_a)}+\beta_5 JulianDate_{(lon_i,lat_r,jul_a)})}\Bigg)^{-1} \end{aligned} \tag{6}$$

is the model prediction of the probability, that the PE at spatio-temporal location (lon_i, lat_r, jul_a) is an EPE

$$P(EPE_{(lon_i,lat_r,jul_a)}|\mathbf{x}_{(lon_i,lat_r,jul_a)},\beta) = p_{(lon_i,lat_r,jul_a)}^{EPE(lon_i,lat_r,jul_a)}(1-p_{(lon_i,lat_r,jul_a)})^{1-EPE_{(lon_i,lat_r,jul_a)}} \tag{7}$$

is the likelihood of a single classification decision

$$\begin{aligned} L(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5) &= \prod_{(lon_i,lat_r,jul_a)\in SP} P(EPE_{(lon_i,lat_r,jul_a)}|\mathbf{x}_{(lon_i,lat_r,jul_a)},\beta) \\ &= \prod_{(lon_i,lat_r,jul_a)\in SP} p_{(lon_i,lat_r,jul_a)}^{EPE_{(lon_i,lat_r,jul_a)}}(1-p_{(lon_i,lat_r,jul_a)})^{1-EPE_{(lon_i,lat_r,jul_a)}} \end{aligned} \tag{8}$$

is the likelihood function or the likelihood of the entire training data, respectively

$$\boldsymbol{\beta}_c = \min_{\boldsymbol{\beta}} \sum_{(lon_i,lat_r,jul_a)\in SP} w_{(lon_i,lat_r,jul_a),c}\left(EPE_{(lon_i,lat_r,jul_a)} - x_{(lon_i,lat_r,jul_a)}^T\beta\right)^2 \tag{9}$$

is the estimate of the $\boldsymbol{\beta}$ through *Iteratively Reweighted Least Squares*

4.3 Generalized *Additive Mixed* Models (GAMM) with a logit link-function and spline-based smooth functions as non-linear special cases of GAMMs

Presently, following the above mentioned steps of generalization and on a conceptual level, the spatio-temporal classification of PE as EPE is arguably already partially modelled: The Logistic Regression Model has a probability value as its response after the superimposition of a threshold-layer, the classification of the PE as EPE can be modeled as depending on a linear predictor consisting out of covariates such as altitude, mean-temperature, spatial location and temporal location, while the linear predictor is supplied to the inverse of the logit link-function. From here, the remaining two requirements, the model's ability to capture spatial, temporal and spatio-temporal dependencies as well as capture non-linear relations between response and predictor, are

introduced in two simultaneous steps: Another generalization, namely the substition of the linear effects through smoothing-terms, as well as the introduciton of random effects. As a consequence, the originally linear predictor is no longer linear and since the smoothing terms are still additively concatenated, ultimately a Generalized Additive Mixed Model with a logit link-function and smoothing terms for effects is built. As this GAMM is able to model the probability of a PE being an EPE as being spatio-temporally dependent on the mentioned covariates, it takes the superimposition of a filtering probability threshold-layer to produce a classifier model that finally meets all the discussed requirements.

5 The Necessity for modelling a *SPATIO-TEMPORAL* classification procedure

Within this thesis, I assert, that the clasifier models to be built have to account for spatial, temporal and spatio-temporal dependencies in the precipitation heights data. To support this assertion, this chapter is set out to demonstrate these spatial and temporal dependencies, where the variation of spatial and temporal covariability will be visualized and analyzed accordingly. Tools, that are used to demonstrate these dependencies are...

6 Appendix

6.1 Different Models at the Core of Classifier Models

6.1.1 Classical Linear Regression Model

$$\begin{aligned}\textbf{Precipitation Height} = \beta_0 + \beta_1 \textbf{Mean Temperature} + \beta_2 \textbf{Altitude} \\ + \beta_3 \textbf{Longitude} + \beta_4 \textbf{Latitude} + \beta_5 \textbf{Julian Date} + \epsilon\end{aligned} \tag{10}$$

---

with assumptions

Linearity:

$y_{(lon_i, lat_r, jul_a)} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{(lon_i, lat_r, jul_a, j)} + \epsilon_{(lon_i, lat_r, jul_a)}$

Independence:

$E(\epsilon_{(lon_i, lat_r, jul_a)}) = 0,\ Var(\epsilon_{(lon_i, lat_r, jul_a)}) = \sigma^2,\ Cov(\epsilon_{(lon_i, lat_r, jul_a)}, \epsilon_{(lon_k, lat_s, jul_b)}) = 0$
$\forall\ (lon_i, lat_r, jul_a) \neq (lon_k, lat_s, jul_b)$

Homoscedasticity:

$Var(\epsilon_{(lon_i, lat_r, jul_a)}) = \sigma^2\ \forall\ (lon_i, lat_r, jul_a) \in\ SP$

Normality:

$\epsilon_{(lon_i, lat_r, jul_a)} \sim N(0, \sigma^2)$

No perfect multicollinearity:

$\text{rank}(X) = p$

No autocorrelation:

$$Cov(\epsilon_{(lon_i, lat_r, jul_a)}, \epsilon_{(lon_k, lat_s, jul_b)}) = 0 \qquad \forall \qquad \begin{pmatrix} (lon_i - lon_k) \\ (lat_r - lat_s) \\ (jul_a - jul_b) \end{pmatrix} \neq \mathbf{0}$$

Declaration of Symbols and Equation Components

Equation 10

Precipitation Height:

Response

β_0:

Intercept parameter.

β_1:

Coefficient for $MeanTemperature$.

β_2:

Coefficient for $Altitude$.

β_3:

Coefficient for $Longitude$.

β_4:

Coefficient for $Latitude$.

β_5:

Coefficient for $JulianDate$.

ϵ:

Error term.

Assumptions

Linearity:

$y_{(lon_i, lat_r, jul_a)}$:

Response.

β_j:

Coefficient for the j-th predictor variable.

$x_{(lon_i, lat_r, jul_a, j)}$:

value of j -th predictor variable at space-time location (lon_i, lat_r, jul_a).

Independence:

$E(\epsilon_{(lon_i, lat_r, jul_a)})$:

Expected value of error term for precipitation height observed at space-time location.

$Var(\epsilon_{(lon_i, lat_r, jul_a)})$:

Variance of error term at space-time location.

$Cov(\epsilon_{(lon_i, lat_r, jul_a)}, \epsilon_{(lon_k, lat_s, jul_b)})$:

Covariance of error terms between two space-time locations.

Homoscedasticity:

σ^2:

Constant variance of error term.

$(lon_i, lat_r, jul_a) \in: SP$:

Space-time locations as elements of the space-time domain derived from the combination of the two-dimensional spatial domain (longitude,latitude) and the temporal domain made up of julian dates.

Normality:

$\epsilon_{(lon_i, lat_r, jul_a)}$:

Error term.

$N(0, \sigma^2)$:

Normal distribution with mean 0 and variance σ^2.

No perfect multicollinearity:

rank(X):

Rank of the design matrix X.

p:

Number of predictor variables.

No autocorrelation:

$Cov(\epsilon_{(lon_i, lat_r, jul_a)}, \epsilon_{(lon_k, lat_s, jul_b)})$:

Covariance of error terms.

$\Big((lon_i - lon_k)\ (lat_r - lat_s)\ (jul_a - jul_b)\Big)$:

Vector of differences in longitude, latitude, and Julian date between two locations and/or time points.

6.1.2 Logistic Regression Model

6.1.3 Generalized Additive Mixed Model

$$\begin{aligned}\text{logit}(p(PE_{lon_i,lat_r,jul_a}\ is\ in\ class\ EPE)) = \beta_0 &+ f_{lon}(lon_{lon_i,lat_r,jul_a}) + f_{lat}(lat_{lon_i,lat_r,jul_a}) \\ &+ f_{jul}(jul_{lon_i,lat_r,jul_a}) + f_{mean\_temp}(mean\_temp_{lon_i,lat_r,jul_a}) \\ &+ f_{altitude}(altitude_{lon_i,lat_r,jul_a}) + b_{lon_i,lat_r,jul_a}\end{aligned} \tag{11}$$

where $logit(p_i) := \ln\left(\frac{p_i}{1-p_i}\right)$ is the logit function of the predicted probability of an event i belonging to the class, b_i is a random effect term to account for unobserved heterogeneity in the data, and f_{lon}, f_{lat}, f_{jul}, f_{mean_temp}, and $f_{altitude}$ are smooth functions of the corresponding predictor variables.

$$\begin{aligned}p_i = \frac{1}{1+\exp(-(\beta_0 + f_{lon}(lon_i) + f_{lat}(lat_r)} \\ + f_{jul}(jul_a) + f_{mean\_temp}(mean\_temp_i) + f_{altitude}(altitude_i))\end{aligned} \tag{12}$$

where $lon_{(lon_i,lat_r,jul_a)}$, $lat_{(lon_i,lat_r,jul_a)}$, $jul_{(lon_i,lat_r,jul_a)}$, $mean\_temp_{(lon_i,lat_r,jul_a)}$, and $altitude_{(lon_i,lat_r,jul_a)}$ are the values of the longitudinal degree, latitudinal degree, Julian date, mean daily temperature, and altitude predictor variables for precipitation even $PE_{(lon_i,lat_r,jul_a)}$, and $flong_degree$ to $faltitude$ are smooth functions of these variables.

To convert this GAMM into a classifier model, we add a threshold layer, which classifies events as belonging to the class if their predicted probability of belonging to the class exceeds the threshold:

The classifier model with the threshold layer can be expressed as:

$$y_i = \begin{cases} 1, & \text{if } p_i > \theta \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

where y_i is the predicted class label for event i, and θ is the chosen threshold value.

6.1.4 Inverse Distance Weighting

Everything is related to everything else, but near things are more related than distant things.

- Waldo Tobler (1970)

The Generalized Additive Mixed Model with a Logit Link function, as expressed in the subsection above, relies on an algorithm to identify the dependence of precipitation height on mean temperature, altitude, spatial and temporal location. This dependence is broadly speaking expressed as the estimated coefficients and respective smoothing terms. As described, it is indeed suitable for picking up on spatial and temporal dependencies due to its unique modelling concept. However, the mathematical foundation of the formulation and estimation of a GAMM is not as easily accessible as another non-statistical, deterministic interpolation method called "Inverse Distance Weighting, which in comparison requires substantially less statistical expertise.
Compared to the aforementioned models, Inverse Distance Weighting does not model the conditional expected value of a variable like precipitation height, nor the conditional probability for a specific observation to be a member of a certain class. It models the variable based on a characteristic of its own univariate distribution: its neighborhood in terms of spatial and temporal distance. For any spatio-temporal location, the existing available observations are evaluated against their relative spatial and temporal distance to the respective spatio-temporal location. The model prediction of the variable for the respective location is then determined as a weighted average of all available observed values minus any existing observations at the location in question.

$$\widehat{rain}_{(lon_i,lat_r,jul_a)} := \sum_{(lon_k,lat_s,jul_b)\in SP} \Big(w_{((lon_k,lat_s,jul_b),(lon_i,lat_r,jul_a))} \cdot rain_{(lon_k,lat_s,jul_b)} \Big) \qquad (14)$$

---

where

$$d^{\alpha}_{((lon_i,lat_r,jul_a),(lon_k,lat_s,jul_b))} := \left(\sqrt{(lon_i - lon_k)^2 + (lat_r - lat_s)^2 + C(jul_a - jul_b)^2}\right)^{\alpha} \tag{15}$$

and

$$w_{((lon_i,lat_r,jul_a),(lon_k,lat_s,jul_b))} := \frac{d^{-\alpha}_{((lon_i,lat_r,jul_a),(lon_k,lat_s,jul_b))}}{\sum_{(lon_k,lat_s,jul_b)\in EP} d^{-\alpha}_{((lon_i,lat_r,jul_a),(lon_k,lat_s,jul_b))}} \tag{16}$$

Declaration of Symbols and Equation Components

$\widehat{\mathbf{rain}}(\mathbf{lon_i}, \mathbf{lat_r}, \mathbf{jul_a})$:

The estimated precipitation height at a specific location and time, denoted by longitude lon_i, latitude lat_r, and Julian day jul_a.

$\sum_{(\mathbf{lon_k},\mathbf{lat_s},\mathbf{jul_b})\in\mathbf{SP}}$:

A summation over all the locations and times in a specific spatio-temporal domain SP or set of all spatio-temporal locations, respectively. The location is denoted by longitude lon_k, latitude lat_s, and Julian day jul_b.

$\left(\mathbf{w}_{((\mathbf{lon_k},\mathbf{lat_s},\mathbf{jul_b}),(\mathbf{lon_i},\mathbf{lat_r},\mathbf{jul_a}))} \cdot \mathbf{rain}_{(\mathbf{lon_k},\mathbf{lat_s},\mathbf{jul_b})}\right)$:

The product of two terms. The first term, $w_{((lon_k,lat_s,jul_b),(lon_i,lat_r,jul_a))}$, is a weight assigned to each location and time in SP based on its distance from the location and time of interest. The second term, $rain_{(lon_k,lat_s,jul_b)}$, is the available observed precipitation height at that location and time.

6.2 Variation of Spatial, Temporal and Spatio-Temporal Covariability

As part of the general argument of this thesis, I have asserted that there are spatial, temporal, and spatio-temporal dependencies within the precipitation height data that should be conceptually accounted for when modelling the classification procedure. In this section, you can find equations on which the calculation of the variation of covariability is based, which is then visualized and analyzed to demonstrate the earlier-mentioned dependencies. Reflecting the different types of spatial, temporal, and spatio-temporal covariability, the equations for the empirical spatial and temporal semivariogram, the spatial and temporal covariance function, and the spatio-temporal covariance function are depicted and described in detail.

In this context, spatial, temporal, and spatio-temporal dependency are understood to be patterns where the similarity between observations decreases as the distance between the observation locations increases. These measures can then calculate the variation in the covariability of the precipitation height data between a pair of spatial, temporal, or spatio-temporal locations as the distance between the paired locations changes. As covariability is commonly used to indicate similarity in data, the variation of covariability as a function of spatial, temporal, or spatio-temporal lags can hence be utilized to assess the prevalence of the various types of dependencies.

6.2.1 Haversine Formula: Calculating the distance between two pairs of angular coordinates

A necessary prerequisite for calculating the variation of covariability is to determine a concept for the various lags first, or in other words, to decide on a way to measure distance. For the spatial distance between two weather stations, the Haversine Formula is used to measure the distance between two points on the Earth's surface based on angular coordinates. The Haversine Formula takes the curvature of the Earth's surface into account and provides an estimate of the so-called great-circle distance between two points. With reference to the quality of measurement concept, it exceeds, for instance, the Euclidean distance, which assumes a flat sphere.

Haversine Formula

$$d \coloneqq 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{17}$$

Declaration of Symbols and Equation Components

d:

Distance between two spatial points in kilometre

r:

Radius of the world in kilometre (6371km)

ϕ_1 and ϕ_2:

Latitudinal Degrees of the two points measured in radians. ϕ is calculated from a decima latidunial degree per $\phi = lat\ \frac{\pi}{180}$, where lat is a latitudinal degree.

λ_1 and λ_2:

Longitudinal Degrees of the two points measured in radians. λ is calculated from a decimal latitudinal degree per $\lambda = lon\ \frac{\pi}{180}$

sin and **cos**:

Trigonometric functions. For a given angle measured in radians, both functions return either the sine or the cosine.

arcsin:

$arcsin(x)$ is the inverse of the sine function. For a given sine x, it returns the sine, whose angle measured (measured in radians) is then x

6.2.2 Empirical Spatial Semivariogram

$$\gamma(d_{spatial}, jul_a) \coloneqq \frac{1}{2|N(d_{spatial})|} \times \left(\sum_{(lon_r, lat_i), (lon_s, lat_k) \in N(d_{spatial})} (rain_{lon_r, lat_i, jul_a} - rain_{lon_s, lat_k, jul_a})^2 \right) \tag{18}$$

Declaration of Symbols and Equation Components

$\gamma(\mathbf{d_{spatial}}, \mathbf{jul_a})$:

The semivariogram function for the precipitation height data over a given interval of spatia distance in kilometre $d_{spatial}$ on a given day jul_a.

$\mathbf{d_{spatial}}$:

An interval of spatial distance between weather stations measured in kilometre within which the precipitation height data are considered to be correlated.

$\mathbf{N}(\mathbf{d_{spatial}})$:

The set of pairs of weather stations, for whom the distance between them lies within the interval $d_{spatial}$.

$|\mathbf{N}(\mathbf{d_{spatial}})|$:

The number of pairs of weather stations in the set $N(d_{spatial})$.

$(\mathbf{lon_r}, \mathbf{lat_i})$ and $(\mathbf{lon_s}, \mathbf{lat_k})$:

Two combinations of longitudinal and latitudinal degree and each denote the spatial location of a weather station

$\mathbf{rain}_{\mathbf{lon_r},\mathbf{lat_i},\mathbf{jul_a}}$ and $\mathbf{rain}_{\mathbf{lon_s},\mathbf{lat_k},\mathbf{jul_a}}$:

Observations of precipitation height at the locations (lon_r, lat_i) and (lon_s, lat_k), respectively, for the Julian day jul_a.

6.2.3 Empirical Temporal Semivariogram

$$\gamma(d_{julian}, (lon_r, lat_i)) \coloneqq \frac{1}{2|N(d_{julian})|} \times \left(\sum_{jul_a, jul_b \in N(d_{julian})} (rain_{lon_r, lat_i, jul_a} - rain_{lon_r, lat_i, jul_b})^2 \right) \tag{19}$$

Declaration of Symbols and Equation Components

$\gamma(\mathbf{d_{julian}}, (\mathbf{lon_r}, \mathbf{lat_i}))$:

The temporal semivariogram function for the precipitation height data at the location (lon_r, lat_i) over a given interval of temporal distances d_{julian}.

$\mathbf{d_{julian}}$:

An Interval of temporal distances within which the precipitation data are considered to be correlated.

$\mathbf{N}(\mathbf{d_{julian}})$:

The set of pairs of Julian days, for whom their distance evaluates to a value within a certain interval d_{julian}.

$|\mathbf{N}(\mathbf{d_{julian}})|$:

The number of pairs of Julian days within the set $N(d_{julian})$.

$\mathbf{jul_a}$ and $\mathbf{jul_b}$:

Are two Julian days

$\mathbf{rain_{lon_r, lat_i, jul_a}}$ and $\mathbf{rain_{lon_r, lat_i, jul_b}}$:

Observations of precipitation height at the location (lon_r, lat_i) and the Julian days jul_a and jul_b, respectively.

6.2.4 Empirical Spatial Covariance Equation

$$\widehat{C}_{rain}^{(d_{julian})}\left(\left(lon_r, lat_i\right), \left(lon_s, lat_k\right)\right) \coloneqq \frac{1}{T - d_{julian}} \times \left(\sum_{jul_a=1+d_{julian}}^{T} \left(rain_{lon_r, lat_i, jul_a} - \widehat{\mu}_{rain,temporal}(lon_r, lat_i)\right) \left(rain_{lon_s, lat_k, jul_a - d_{julian}} - \widehat{\mu}_{rain,temporal}(lon_s, lat_k)\right) \right) \tag{20}$$

Declaration of Symbols and Equation Components

$\widehat{\mathbf{C}}_{\mathbf{rain}}^{(\mathbf{d_{julian}})}$:

The estimated spatial covariance of the precipitation data for a given temporal distance d_{julian}.

$(\mathbf{lon_r}, \mathbf{lat_i})$ and $(\mathbf{lon_s}, \mathbf{lat_k})$:

Two pairs of longitude and latitude coordinates.

$\mathbf{T}$:

The number of time lags used in the estimation of the covariance.

$\mathbf{d_{julian}}$:

Is the temporal distance between the two sets of precipitation data being compared.

$\mathbf{rain_{lon_r, lat_i, jul_a}}$:

Is the observation of precipitation height at the spatial location (lon_r, lat_i) and time jul_a.

$\widehat{\mu}_{\mathbf{rain,spatial}}(\mathbf{lon_r}, \mathbf{lat_i})$ and $\widehat{\mu}_{\mathbf{rain,spatial}}(\mathbf{lon_s}, \mathbf{lat_k})$:

The estimated means of the precipitation data at locations $(\mathbf{lon_r}, \mathbf{lat_i})$ and $(\mathbf{lon_s}, \mathbf{lat_k})$, respectively.

6.2.5 Empirical Temporal Covariance Equation

$$\widehat{C}_{rain}^{(d_{spatial})}\left(jul_a, jul_b\right) \coloneqq \frac{1}{|N(d_{spatial})|} \times \left(\sum_{(lon_r, lat_i),(lon_s, lat_k) \in N(d_{spatial})} \left(rain_{lon_r, lat_i, jul_a} - \widehat{\mu}_{rain,spatial}(jul_a)\right) \left(rain_{lon_s, lat_k, jul_b} - \widehat{\mu}_{rain,spatial}(jul_b)\right) \right) \tag{21}$$

Declaration of Symbols and Equation Components

$\widehat{\mathbf{C}}\mathbf{rain}^{(\mathbf{dspatial})}$:

The estimated covariance of the precipitation data for a given spatial distance $d_{spatial}$.

$\mathbf{jul_a}$ and $\mathbf{jul_b}$:

Julian days that are a certain temporal distance apart.

$\mathbf{N}(\mathbf{d_{spatial}})$:

A set of pairs of spatial locations, for whom the spatial distance between the paired locations is within a certain interval of spatial distance $d_{spatial}$.

$|\mathbf{N}(\mathbf{d_{spatial}})|$:

The number of pairs spatial locations / weather stations, for whom the spatial distance lies within the interval of spatial distance $d_{spatial}$.

$(\mathbf{lon_r}, \mathbf{lat_i})$:

A pair of longitude and latitude coordinates.

$\mathbf{rain_{lon_r, lat_i, jul_a}}$:

The precipitation height value at the location (lon_r, lat_i) and time jul_a.

$\widehat{\mu}_{\mathbf{rain,temporal}}(\mathbf{jul_a})$ and $\widehat{\mu}_{\mathbf{rain,temporal}}(\mathbf{jul_b})$:

The estimated means of the precipitation height data for times jul_a and jul_b, respectively.

6.2.6 Empirical Spatio-Temporal Covariance Equation

$$\widehat{C}_{rain}(d_{spatial}, d_{julian}) \coloneqq \frac{1}{|N(d_{spatial})|}\frac{1}{|N(d_{julian})|} \times \left(\sum_{(lon_r,lat_i),(lon_s,lat_k) \in N(d_{spatial})} \; \sum_{jul_a,jul_b \in N(d_{julian})} (rain_{lon_r,lat_i,jul_a} - \widehat{\mu}_{jul_a})\,(rain_{lon_s,lat_k,jul_b} - \widehat{\mu}_{jul_b}) \right) \tag{22}$$

Declaration of Symbols and Equation Components

$\widehat{\mathbf{C}}_{\mathbf{rain}}(\mathbf{d}_{\mathbf{spatial}}, \mathbf{d}_{\mathbf{julian}})$:

The covariance of the precipitation height data for a given spatial distance in kilometre $d_{spatial}$ and a given temporal distance d_{julian}.

$\mathbf{d}_{\mathbf{spatial}}$:

The interval of spatial distances between two weather stations denoted each as a pair of angular coordinates, within which the precipitation heights are considered to be correlated.

$\mathbf{d}_{\mathbf{julian}}$:

The interval of temporal distances between jul_a and jul_b, within which the precipitation heights are considered to be correlated

$\mathbf{jul_a}$ and $\mathbf{jul_b}$:

Julian days or the days elapsed since the earliest observation within the precipitation data, respectively.

$\mathbf{N}(\mathbf{d}_{\mathbf{spatial}})$:

A set of pairs of weather stations, for whom the spatial distance within the pair evaluates to a value within the interval $d_{spatial}$.

$\mathbf{N(d_{julian})}$:

A set of pairs of weather stations, for whom the temporal distance within the pair evaluates to a value within the interval d_{julian}.

$|\mathbf{N(d_{spatial})}|$:

The number of pairs weather stations within the set $d_{spatial}$.

$|\mathbf{N(d_{julian})}|$:

The number of pairs of temporal locations of the weather stations that are within the set d_{julian}).

$\sum_{\mathbf{(lon_r,lat_i)},\mathbf{(lon_s,lat_k)} \in \mathbf{N(d_{spatial})}} \sum_{\mathbf{jul_a},\mathbf{jul_b} \in \mathbf{N(d_{julian})}}$:

A double summation over pairs of observations of the precipitation height made at a pair of weather stations spatio-temporally located, such that the pair belongs to $d_{spatial}$ spatially and d_{julian} temporally.

$\mathbf{rain_{lon_r,lat_i,jul_a}}$:

The precipitation height value at the location (lon_r, lat_i) and time jul_a.

$\widehat{\mu}_{\mathbf{rain,temporal}}(\mathbf{jul_a})$ and $\widehat{\mu}_{\mathbf{rain,temporal}}(\mathbf{jul_b})$:

The spatial mean of the precipitation height data at time jul_a and jul_b.

$(\mathbf{rain_{lon_r,lat_i,jul_a}} - \widehat{\mu}_{\mathbf{rain,temporal}}(\mathbf{jul_a}))$:

The deviation of the precipitation height data from their mean at location (lon_r, lat_i) and time jul_a.

$(\mathbf{rain_{lon_s,lat_k,jul_b}} - \widehat{\mu}_{\mathbf{rain,temporal}}(\mathbf{jul_b}))$:

The deviation of the precipitation height data from their mean at location (lon_s, lat_k) and time jul_b.

The product of the two deviations:

$$(rain_{lon_r,lat_i,jul_a} - \widehat{\mu}_{rain,temporal}(jul_a))(rain_{lon_s,lat_k,jul_b} - \widehat{\mu}_{rain,temporal}(jul_b)).$$

$\frac{\mathbf{1}}{\mathbf{2}|\mathbf{N}(\mathbf{d_{spatial}})|} \frac{\mathbf{1}}{\mathbf{2}|\mathbf{N}(\mathbf{d_{julian}})|}$:

A normalization factor that scales the covariance to account for the number of pairs of weather stations, for whom the distance in kilometre evaluates to a value within $d_{spatial}$, while their temporal distance evaluates to value within the interval $d_{temporal}$ being used in the calculation.

7 Session Info