Master's Thesis

Predicting Droughts in the Amazon Basin based on Global Sea Surface Temperatures

Department of Statistics
Ludwig-Maximilians-Universität München

Dario Lepke

Munich, August 9th, 2022

Submitted in partial fulfillment of the requirements for the degree of M. Sc.
Supervised by Dr. Fabian Scheipl (LMU) and Dr. Niklas Boers (PIK)

Contents

Introduction **9**

1 Related work **11**

2 EDA **13**

2.1 EDA precipitation 13

2.2 Glyph plots 16

2.3 EDA SST 16

3 Correlation analysis **17**

3.1 Correlation of Sea Surface Temperature and Precipitation 17

3.2 Summary 17

4 Clustering **19**

4.1 Main Idea Clustering 19

4.2 Clustering Methods 20

4.3 Clustering results 20

4.4 Analyse clustering results 20

5 LASSO Regression **21**

5.1 Introduction 21

5.2 Implementation 21

5.3 TODO here 21

5.4 Results 21

5.5 Lasso summary 21

6 The fused lasso **23**

6.1 Introduction 23

6.2 Implementation 23

6.3 Model evaluation 23

6.4 Graph structure 23

6.5 Results 23
6.6 Summary 23

List of Figures

2.1 Location of the area under study. The central amazon basin (CAB) spanning across 0,-10 latitude and -70,-55 longitude 13

2.2 Density of the raw precipitation values without time or spatial dependency. 14

2.3 Mean and standard deviation at each location. The standard deviation was computed over the whole time period. The white line on the scale at the side of the plots indicates the mean of the respective quantity 14

2.4 Mean precipitation values at each location, shown for the different months of a year separately. The white line on the scale at the side of the plots indicates the mean of the respective quantity 15

List of Tables

Introduction

Placeholder

Chapter 1

Related work

Placeholder

Chapter 2

EDA

In this chapter, we will explore the values of the precipitation and SST data for the common observation period from 1981 until 2016. We analyze the data from three perspectives. Firstly the raw values without time or spatial dependency, second the mean and standard deviation for each spatial grid cell for the whole time series and then the mean and standard deviation for each grid cell but for each month of the year separately.

2.1 EDA precipitation

Here we study the time series of precipitation in the Central Amazon Basin. The CHIRPS data set contains the precipitation data, created from in-situ and satellite measurements (Funk et al. (2015)). It can be downloaded for example, from here. It contains observations from 1981 to 2016 and comes on a high resolution of 0.05 grid, which we aggregate to a 0.5 grid.

In 2.1 we show the area of the Central Amazon Basin that is object of our study.

Figure 2.1: Location of the area under study. The central amazon basin (CAB) spanning across 0,-10 latitude and -70,-55 longitude

Firstly we plot precipitation values in general in Figure 2.2 Its form is a uni-modal, right-skewed density. The values range from 0 up to 769, but only few observations take these high values, forming a large tail. This might be a indication for large outliers in the data or due to some locations with very high precipitation values in general.

As we can see in Figure 2.3 most locations have a mean precipitation of around 200 mm/month, over the whole time series. Regionally in the "upper left" corner of the Amazon Basin, mean precipitation is higher or equal to the mean. The reference point for "higher" is the mean of the location means. This

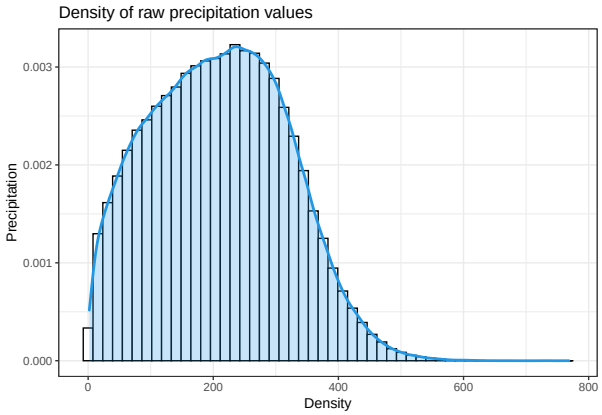


Figure 2.2: Density of the raw precipitation values without time or spatial dependency.

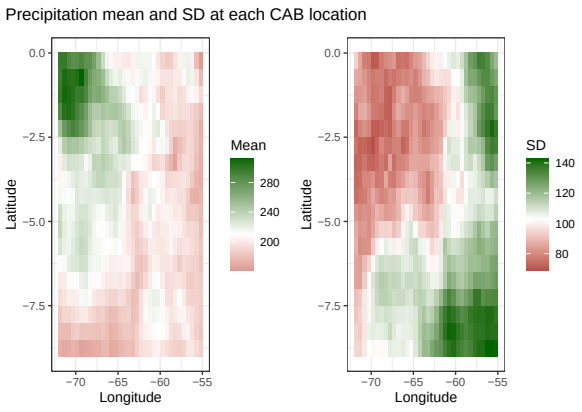


Figure 2.3: Mean and standard deviation at each location. The standard deviation was computed over the whole time period. The white line on the scale at the side of the plots indicates the mean of the respective quantity

region seems to be more or less spatially consistent. The rest of the region with lower mean precipitation has also some small areas where precipitation is again a little bit higher. For example in the upper right corner and on the bottom, right of the middle. For the standard deviation we also see regional patterns. These patterns overlap with the regions of the mean but their magnitude is flipped. Meaning, in the upper left where we observe larger mean values we generally observe lower standard deviation and in the lower and upper right corners, higher standard deviations. We see spatial patterns of the mean evolving over time in (2.4). For example: From May until August there is a spatial separation in two parts that dissolves in September. As expected there is a large seasonal component regarding the means. For the standard deviation we see as well large differences in values during different months of the year.

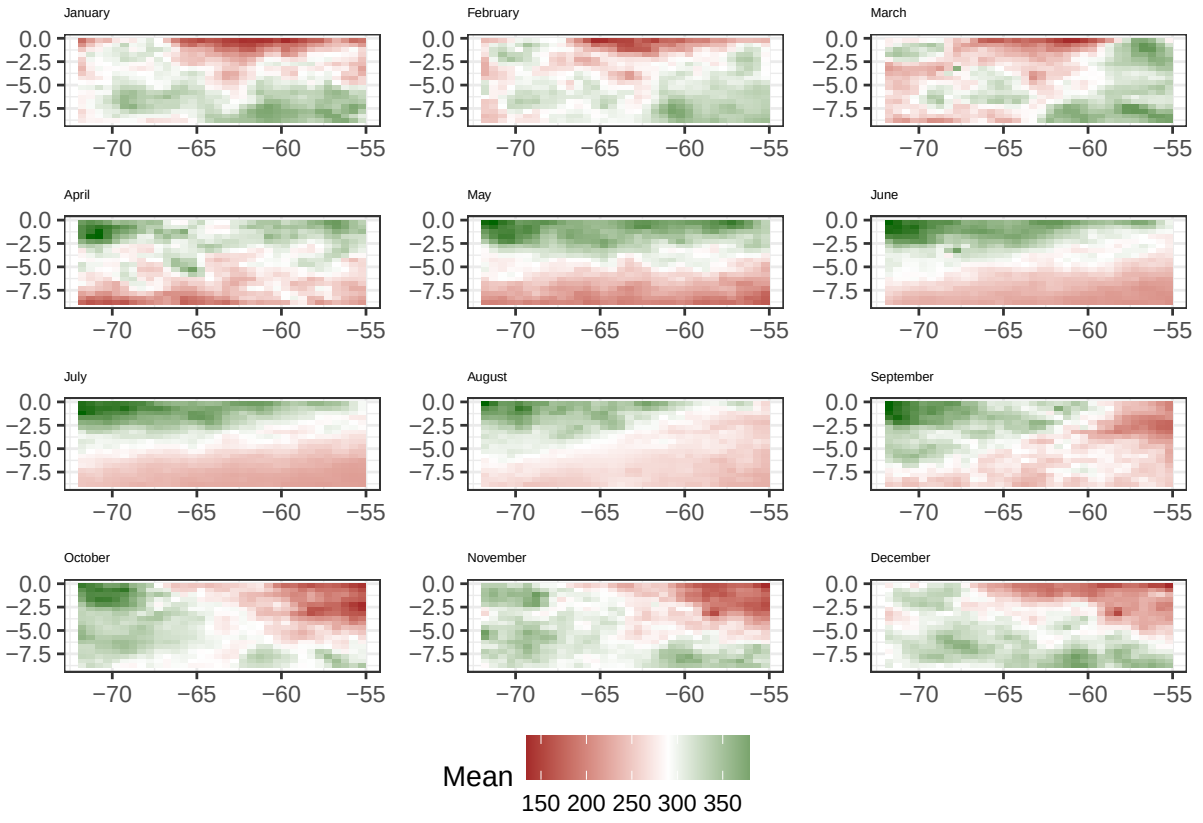
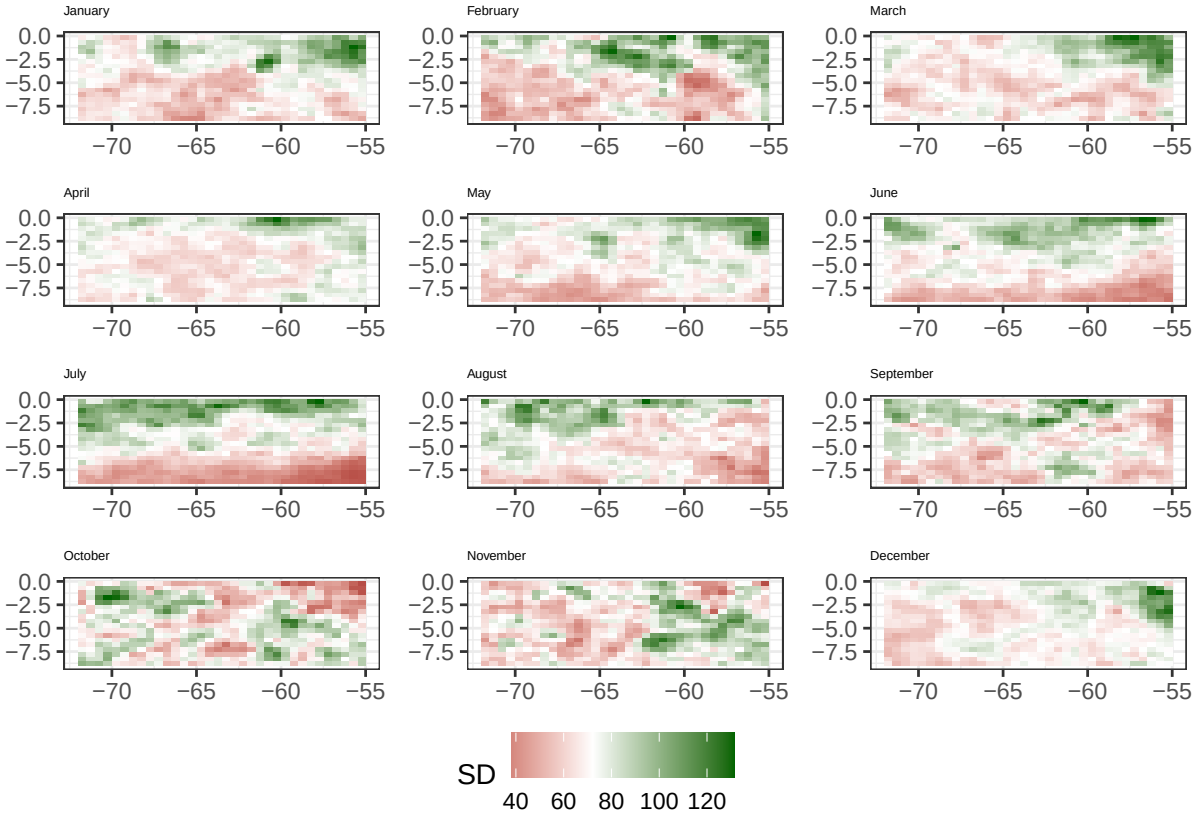


Figure 2.4: Mean precipitation values at each location, shown for the different months of a year separately. The white line on the scale at the side of the plots indicates the mean of the respective quantity



2.2 Glyph plots

2.3 EDA SST

Chapter 3

Correlation analysis

Placeholder

3.1 Correlation of Sea Surface Temperature and Precipitation

3.1.1 Original Data

3.1.1.1 Timelag 0

3.1.1.2 Timelag 3

3.1.1.3 Timelag 6

3.1.1.4 Timelag 12

3.1.2 De-seasonalized Data

3.1.2.1 Timelag 0

3.1.2.2 Timelag 3

3.1.2.3 Timelag 6

3.1.2.4 Timelag 12

3.2 Summary

3.2.1 Original Data

3.2.2 De-seasonalized Data

Chapter 4

Clustering

In this chapter we will first summarize the main ideas of clustering and then apply it to the precipitation data. If not indicated otherwise the information is taken from Elements of Statistical Learning.

4.1 Main Idea Clustering

We can describe an object by a set of measurements or its similarity to other objects. Using this similarity we can put a collection of objects into subgroups or clusters. The objects in the subgroups should then be more similar to one another than to objects of different subgroups. This means inside the clusters we aim for homogeneity and for observations of different clusters for heterogeneity. With the clustering analysis applied to the precipitation data we want to study if there are distinct groups (regions) apparent in the CAB. So that if we later apply the regression models we predict the precipitation for each group and not for the whole region.

To explore the grouping in the data we need a measure of (dis)similarity. This measure is central and depends on subject matter considerations. We construct the dissimilarities based on the measurements taken for each month. We interpret this as a multivariate analysis where, each month is one variable. So given the area in the CAB (resolution 5°x5°), we have 612 cells and 432 months, resulting in a 612×432 data matrix. we want to cluster cells into homogen groups.

4.2 Clustering Methods

4.2.1 k-means

4.2.2 k-means characteristics

4.2.3 K-medoids

4.2.3.1 K-medoids characteristics

4.2.4 PCA

4.2.5 Gap statistic

4.3 Clustering results

4.3.1 k-means and PAM gap statistics without PCA

4.3.1.1 Scree plot

4.3.1.2 k-means and PAM gap statistics after applying PCA

4.3.2 Summary

4.4 Analyse clustering results

Chapter 5

LASSO Regression

Placeholder

5.1 Introduction

5.2 Implementation

5.3 TODO here

5.4 Results

5.4.1 Lasso

5.4.2 standardized lasso

5.4.3 deseas lasso

5.4.4 diff1 lasso

5.4.5 Lasso on clustered precipitation

5.4.5.1 Cluster 1

5.4.5.2 Cluster 2

5.4.5.3 Cluster 3

5.4.5.4 Cluster 4

5.4.5.5 Cluster 5

5.4.5.6 Cluster Summary

5.5 Lasso summary

Chapter 6

The fused lasso

Placeholder

6.1 Introduction

6.2 Implementation

6.3 Model evaluation

6.4 Graph structure

6.5 Results

6.5.1 Fused lasso without clusters

6.6 Summary

6.6.1 Fused Lasso with clusters

Funk, Chris, Pete Peterson, Martin Landsfeld, Diego Pedreros, James Verdin, Shraddhanand Shukla, Gregory Husak, et al. 2015. "The Climate Hazards Infrared Precipitation with Stations-a New Environmental Record for Monitoring Extremes." *Scientific Data* 2 (1): 1–21.