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

Dario Lepke

# Contents

# Introduction

Placeholder

# Chapter 1

# Related work

Placeholder

# Chapter 2

# EDA

Placeholder

## 2.1 EDA precipitation

## 2.2 Glyph plots

This section provides a graphical presentation of the precipitation data known as glyph plots. The idea of glyph maps, its application and general implementation that were used in this section are taken from @**??**wickham2012glyph. Glyph maps use a small icon or *glyph* to show multiple values at each location. In our case, we show a complete time series at each location instead of just single values. Different techniques can then be used compare the time series between all locations or their individual shape on a local scale. We will show seasonal, de-seasonalized, and de-seasonalized data on a local scale. Seasonal time series are computed by computing the averages of each month on each location. Each seasonal time series therefore has only 12 values and can be plotted without smoothing. The de-seasonalized time series are computed by omitting the seasonal effects on each time series for the *complete* observation period and therefore has to be smoothed to be visually inspectable. The de-seasonalized time series then can be used to compare the time series for each location on a common or local scale. On the common scale all values are displayed on the same axis range, while on the local scale the axis are changed so that their ranges refer to the range on the respective location. Rescaling is done as follows

$$x_{rescaled} = \frac{x - \min(x)}{\max(x) - \min(x)}.$$

This will help us to see the changes in value at each location *relative* to the range of the values at the same location. But this also means that interpreting these plots has to be done carefully because, in this form of display, large difference might actually refer to only small changes in absolute values. It can be due to the small range of values at that location in general, that these changes seem to be large. To aid the interpretation of these plots we can use color shadings to draw attention to areas in which ranges are large, meaning larger differences in their relative values also point to larger differences in their absolute values (i.e unscaled values, values on the global scale). Therefore locations with large ranges are shaded in lighter colors and smaller ranges are shaded in dark color, to make the lighter shaded areas more easily visible.

To improve readability of glyph maps, one can also add boxed for each glyph as well as reference lines for global means. This way the trajectory of the glyphs can be viewed in comparison to the overall mean directly.

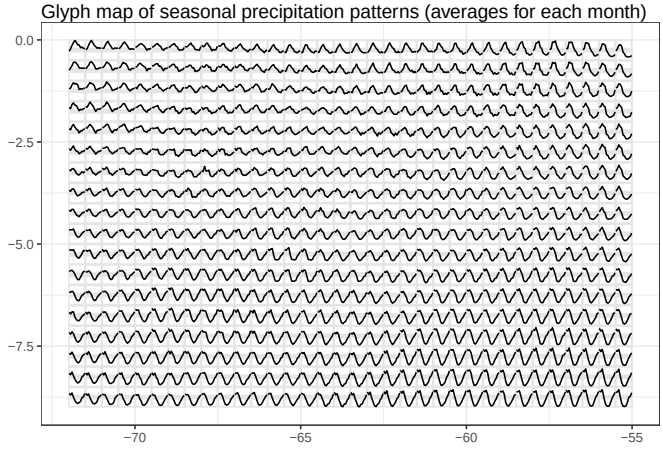


Figure 2.1: Glyph map of seasonal precipitaton pattern. Each location is presented by a time series. The time series are seperated by boxes. The gray reference lines inside the boxes show the mid-range for easier comparison.

The above figure is a glyph-map of seasonal precipitation patterns (averages for each month) in the Central Amazon Basin. The gray reference lines show the mid-range for easier comparison of the patterns. We see differences in the seasonal patterns across the map. In the upper left for example, the seasonal patterns stay above mid-range while on the bottom-left they have values clearly towards the low end of the range. Also some areas have multimodal patterns. The patterns differ in range and month of maximum and minimum precipitation.

This plot shows the smoothed de-seasonalized monthly precipitation, after global scaling. The same position within each cell corresponds to the same value in all locations. Some areas have almost a linear course, increasing, decreasing or constant. Others show a more “wiggly” courses. As overall pattern we can see that the forms of the patterns have a spatial connection,

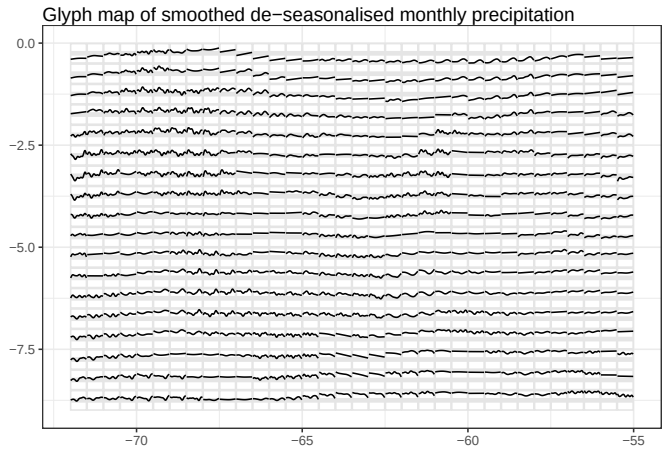


Figure 2.2: Glyph map of de-seasonalised and smoothed precipitation. Each location is presented by a time series. The time series are seperated by boxes. The gray reference lines inside the boxes show the mid-range for easier comparison. The time series are scaled globally, same positions inside the cells correspond to the same values in all locations.

patterns are close to similar patterns, at the same latitude. Also regarding latitude the closer to the equator the less precipitation.

Now we inspect the glyph-map with de-seasonalized locally scaled values. This form of scaling emphasizes the individual shapes. Because of the applied scaling, big patterns may be just be tiny effects. Therefore colors are added according to range. Areas with lighter color have larger ranges than darker areas. The areas with steep linear increases and decreases have smaller ranges than or example the areas below -2.5 latitude in the left.

The results of the precipitation glyphs indicate that the CAB might be separable in different regions. If we can find a way to quantify the differences in these regions and separate them into clusters, we could then apply our regression models to each of these clusters and eventually improve model accuracy on each region as compared to the complete are on average. Therefore in a later section we will discuss and apply clustering algorithms to the precipitation data. But for now we will have a look at the SST data.

## 2.3 EDA SST

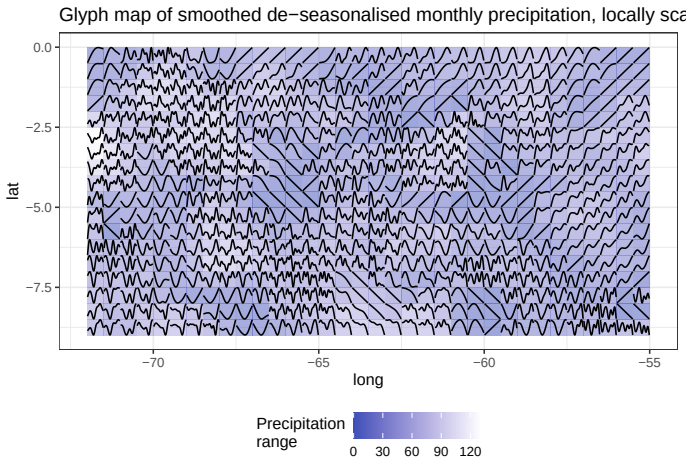


Figure 2.3: Glyph map of de-seasonalised and smoothed precipitation. The time series are scaled locally, ranges are not the same in all cells. The different ranges are given in color shades, where lighter shading indicates a larger range and darker shades smaller ranges.

# Chapter 3

# Correlation analysis

Placeholder

## 3.1 Short Recap

## 3.2 Correlation of Sea Surface Temperature and Precipitation

### 3.2.1 Original Data

#### 3.2.1.1 Timelag 0

#### 3.2.1.2 Timelag 3

#### 3.2.1.3 Timelag 6

#### 3.2.1.4 Timelag 12

### 3.2.2 Deseasonalised Data

#### 3.2.2.1 Timelag 0

#### 3.2.2.2 Timelag 3

#### 3.2.2.3 Timelag 6

#### 3.2.2.4 Timelag 12

## 3.3 Summary

### 3.3.1 Original Data

### 3.3.2 Deseasonalised 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 $612 \times 432$ data matrix. we want to cluster cells into homogen groups.

## 4.2 Clustering Methods

### 4.2.1 K-means

### 4.2.2 Kmeans characteristics

### 4.2.3 K-medoids

#### 4.2.3.1 K-medoids characteristics

### 4.2.4 PCA

### 4.2.5 Gap statistic

## 4.3 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

# Chapter 6

# The fused lasso

Placeholder

## 6.1 Introduction

## 6.2 Implementation