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 . . . 13

2.3 EDA SST . . . 13

3 Correlation analysis **15**

3.1 Short Recap . . . 15

3.2 Correlation of Sea Surface Temperature and Precipitation . . . 15

3.3 Summary . . . 15

4 Clustering **17**

4.1 Main Idea Clustering . . . 17

4.2 Clustering Methods . . . 18

4.3 Clustering results . . . 18

4.4 Analyse clustering results . . . 18

5 LASSO Regression **19**

5.1 Introduction . . . 19

5.2 Implementation . . . 19

5.3 TODO here . . . 19

5.4 Results . . . 19

6 The fused lasso **27**

6.1 Introduction . . . 27

6.2 Implementation . . . 27

List of Figures

5.1 Mean squared error of the 5-fold blocked cross validation for a range of lambda values on the log scale. The points in the middle represent the average MSE for the respective lambda, the errorbars give the MSE +/- one standard deviation. The dotted line shows the lambda for which minimum MSE was obtained. 21

List of Tables

Introduction

Placeholder

Chapter 1

Related work

Placeholder

Chapter 2

EDA

Placeholder

2.1 EDA precipitation

2.2 Glyph plots

2.3 EDA SST

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 612×432 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

```
library(patchwork)
library(ggpubr)
```

```
## Lade nötiges Paket: ggplot2
```

```
library(raster)
```

```
## Lade nötiges Paket: sp
```

```
library(glmnet)
```

```
## Lade nötiges Paket: Matrix
```

```
## Loaded glmnet 4.1-4
```

```
library(Hmisc)
```

```
## Lade nötiges Paket: lattice
```

```
## Lade nötiges Paket: survival
```

```
## Lade nötiges Paket: Formula
```

```
##
## Attache Paket: 'Hmisc'
```

```
## Die folgenden Objekte sind maskiert von 'package:raster':
##
##     mask, zoom
```

```
## Die folgenden Objekte sind maskiert von 'package:base':
##
##     format.pval, units
```

```
source("../code/R/helper-functions.R")
```

```
path_to_model_folder <- "../results/CV-lasso/test-lasso-stand/"
```

```
err_mat <- readRDS(paste0(path_to_model_folder, "/err-mat.rds"))
lambdas <- readRDS(paste0(path_to_model_folder, "/lambda-vec.rds"))
wm <- which.min(apply(err_mat, 1, mean))
full_model <- readRDS(paste0(path_to_model_folder, "full-model.rds"))
intercept <- round(full_model$a0[wm],2)
lambda <- round(lambdas[wm],2)
rm(full_model)
```

```
err_bars_plot <- readRDS(paste0(path_to_model_folder, "/err-mat-plots/err-bars-plot.rds"))
err_bars_plot
```

```
# a <- ggplot_build(err_bars_plot)
```

The learning curve in 5.1 shows some extreme behavior for λ values around 0 on the logarithmic scale. For lower values the standard deviation of the MSE get extremely large. Inspecting the MSE lines sperately for each fold gives more insight.

```
p1 <- readRDS(paste0(path_to_model_folder, "/err-mat-plots/err-plot-fold-1.rds"))
p2 <- readRDS(paste0(path_to_model_folder, "/err-mat-plots/err-plot-fold-2.rds"))
p3 <- readRDS(paste0(path_to_model_folder, "/err-mat-plots/err-plot-fold-3.rds"))
p4 <- readRDS(paste0(path_to_model_folder, "/err-mat-plots/err-plot-fold-4.rds"))
p5 <- readRDS(paste0(path_to_model_folder, "/err-mat-plots/err-plot-fold-5.rds"))

p1 + p2 + p3 + p4 + p5
```

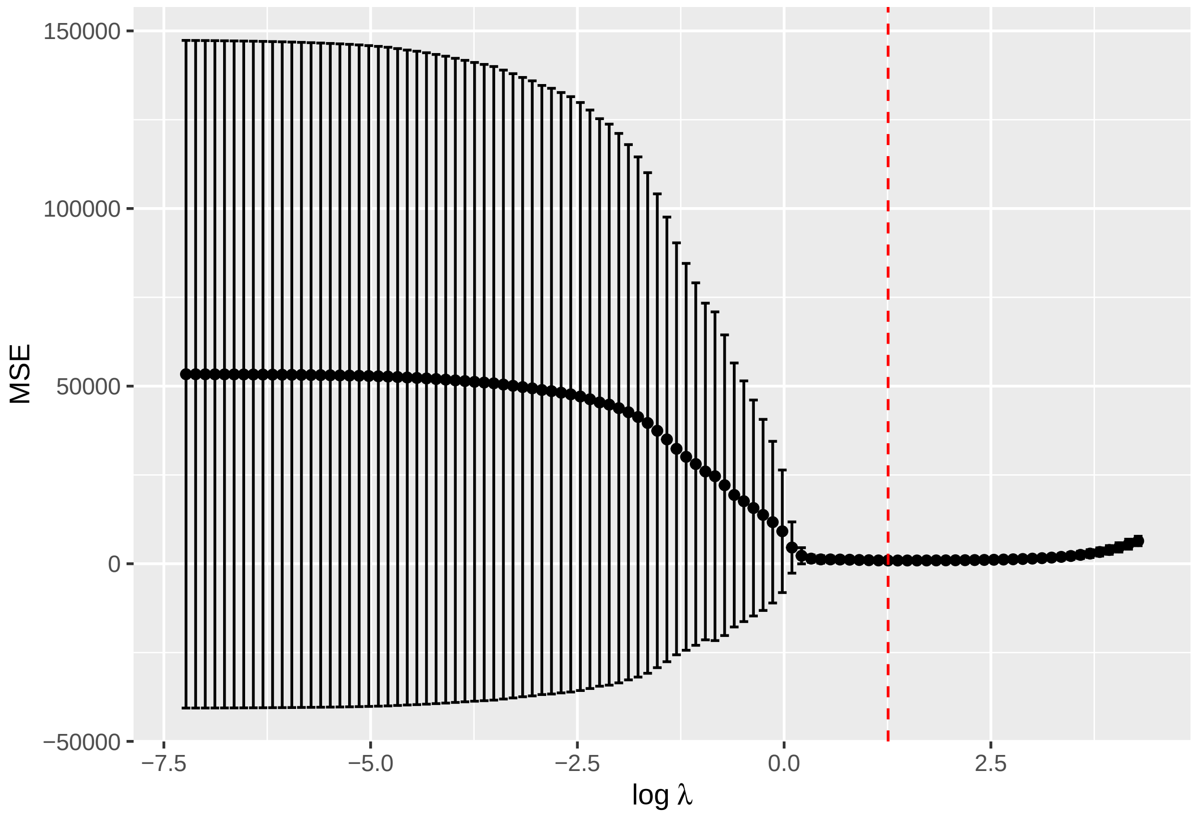
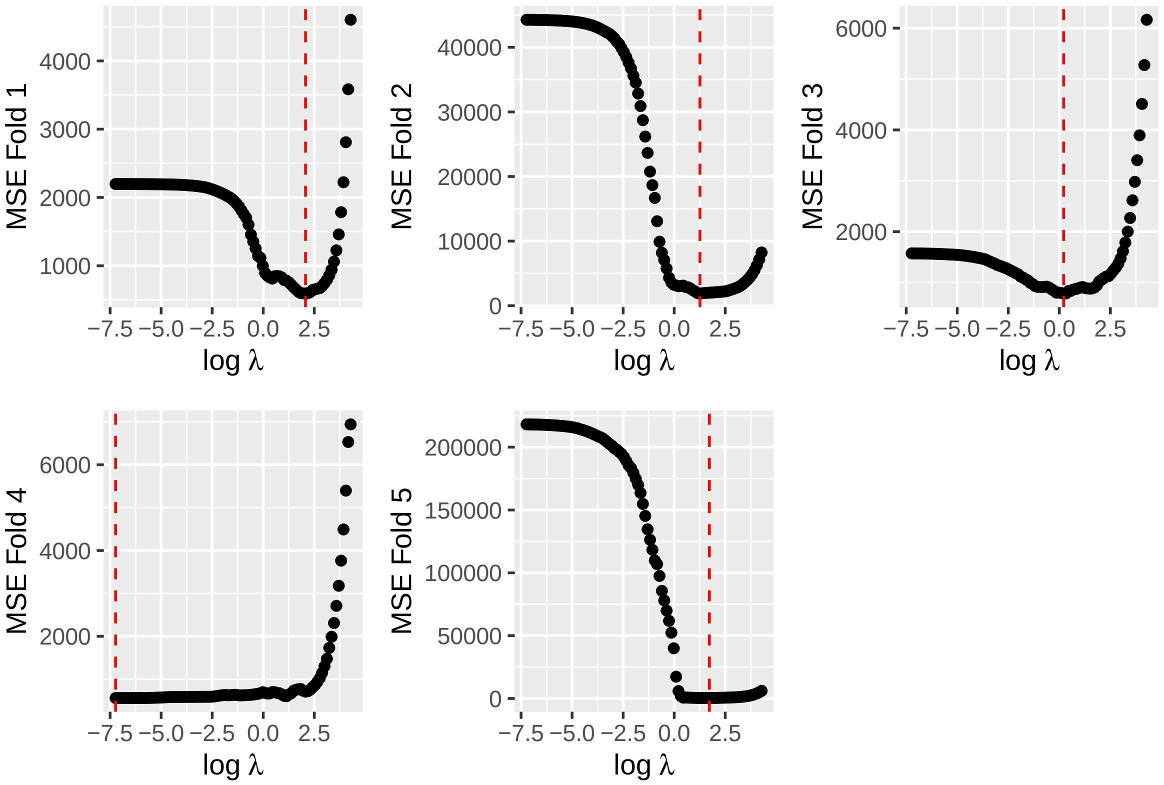


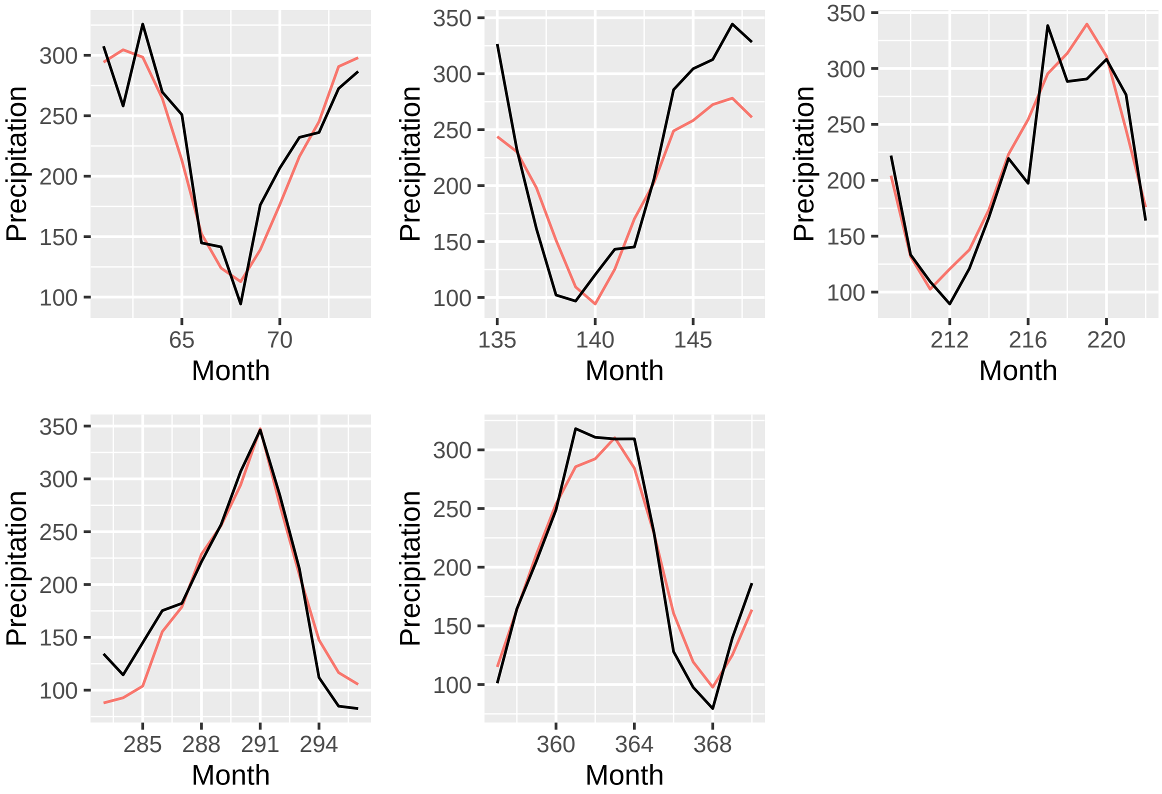
Figure 5.1: Mean squared error of the 5-fold blocked cross validation for a range of lambda values on the log scale. The points in the middle represent the average MSE for the respective lambda, the errorbars give the MSE +/- one standard deviation. The dotted line shows the lambda for which minimum MSE was obtained.



The folds that drive the large MSE standard deviation are fold 4 and 5 (see **??**). The range of the best λ value chosen is larger than for the lasso without standardization. Fold 4 chooses very low regularization as optimal.

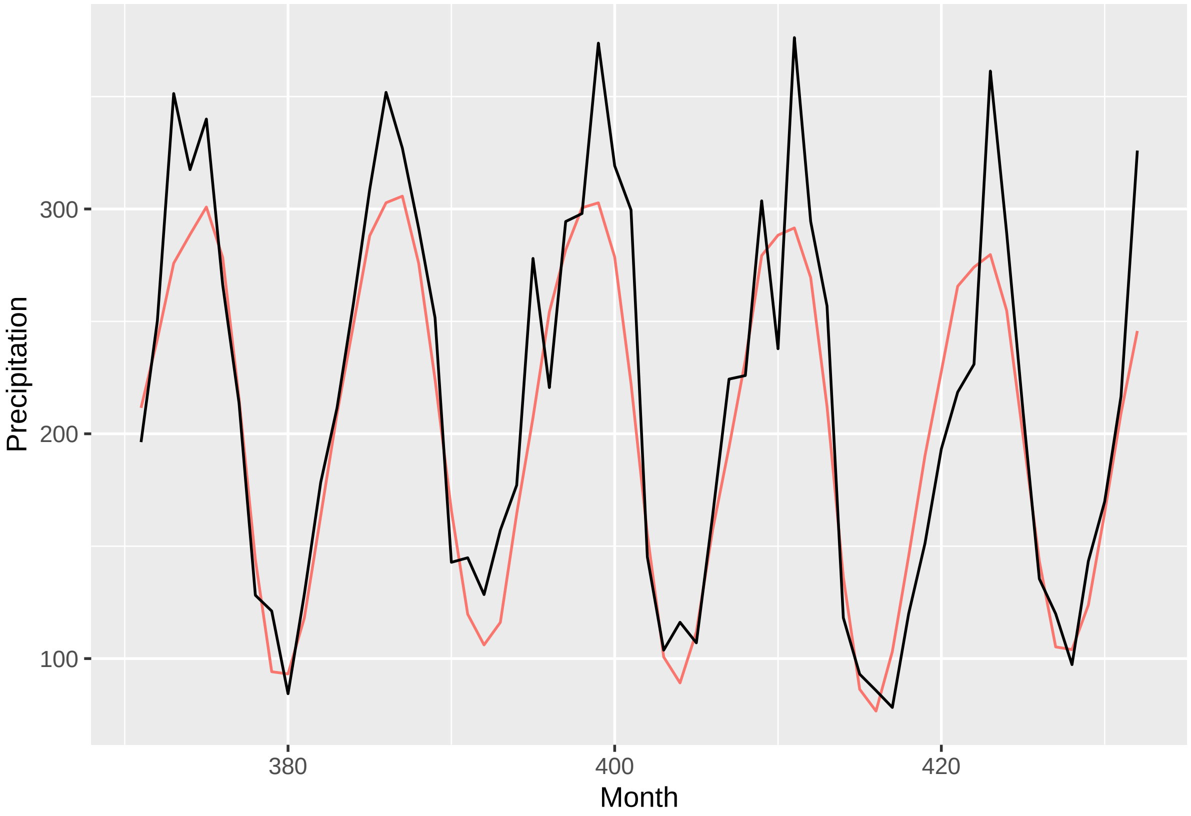
```r
pred_plot_1 <- readRDS(paste0(path_to_model_folder, "/pred-plots/pred-plot-fold-1.rds"))
mse_1 <- get_mse_from_pred_plot(pred_plot_1)

pred_plot_2 <- readRDS(paste0(path_to_model_folder, "/pred-plots/pred-plot-fold-2.rds"))
mse_2 <- get_mse_from_pred_plot(pred_plot_2)


pred_plot_3 <- readRDS(paste0(path_to_model_folder, "/pred-plots/pred-plot-fold-3.rds"))
mse_3 <- get_mse_from_pred_plot(pred_plot_3)

pred_plot_4 <- readRDS(paste0(path_to_model_folder, "/pred-plots/pred-plot-fold-4.rds"))
mse_4 <- get_mse_from_pred_plot(pred_plot_4)

pred_plot_5 <- readRDS(paste0(path_to_model_folder, "/pred-plots/pred-plot-fold-5.rds"))
mse_5 <- get_mse_from_pred_plot(pred_plot_5)

pred_plot_list <- list(pred_plot_1,pred_plot_2,pred_plot_3,pred_plot_4,pred_plot_5)
```

```r
pred_plot_1 + pred_plot_2 + pred_plot_3 + pred_plot_4 +
  pred_plot_5
```



The predictions on the test set are similar to those from the normal lasso, in fold 4 the peak is predicted exactly (**??**). The Model from fold 2 underestimates more than the respective fold model from the lasso without standardization (see **??**).

```
full_preds <- readRDS(paste0(path_to_model_folder, "/pred-plots/pred-plot-full.rds"))
full_preds
```

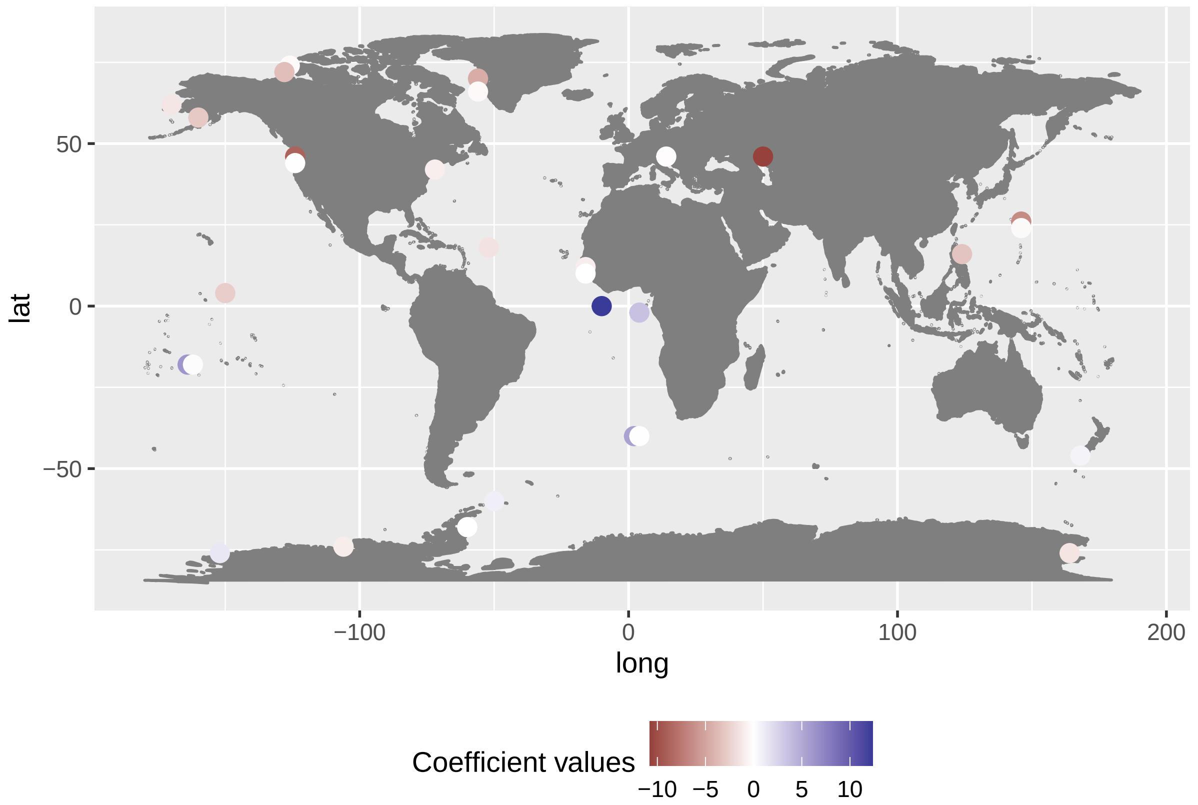


```r
mse_full <- get_mse_from_pred_plot(full_preds)
mse_full
```

```
## [1] 1214.489
```

The predictions from the full model of the standardized lasso show the same behavior as the ones found in the lasso without standardization. Again seasonal patterns are predicted well but largest values are underestimated (@ref:coef-plot-full-lasso-stand).

```r
coef_full <- readRDS(paste0(path_to_model_folder,
                     "coef-plots/coef-plot-full.rds"))
coef_full + theme(legend.position = "bottom")
```



@ref:coef-plot-full-lasso-stand shows now the coefficient plot for the standardized lasso, the coefficient values are given on the standardized scale. The range of coefficient values is more symmetric around the 0 than for first lasso model and the locations differ as well. For example the two locations at pacific coast of South America are not included in the model here (compare @ref:coef-plot-full-lasso-og).

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 Summary

Chapter 6

The fused lasso

Placeholder

6.1 Introduction

6.2 Implementation