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

List of Figures

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 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

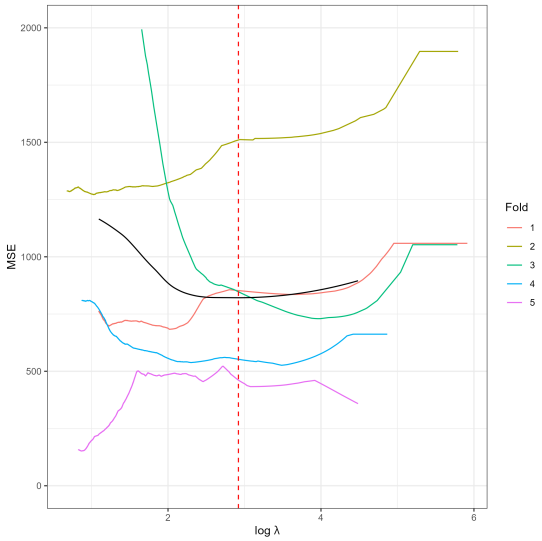
6.3 Model evaluation

```
knitr::opts_chunk$set(echo=FALSE, warning = FALSE, message = FALSE, out.width = '50%', fig.align = 'cen
options(knitr.duplicate.label = "allow")
```

In general we use the same evaluation methods for the fused lasso that we used for the lasso models. We define 5 folds with train and test, search for an optimal regularization value (λ_{min}). Then we fit the model on the complete train and test data with λ_{min} and report the MSE on the evaluation set. But the difference for the lasso and the fused lasso is that we can define a λ vector that we want to search for solutions for the lasso *before* starting the CV, in the fused lasso this is not possible. In the *fusedlasso* function we can define the number of steps the model should take and a λ that defines the end of the path (we will call it λ_{end} here). The model terminates therefore if either the maximum number of steps or the defined minimum λ_{end} is reached (default is 0). This means that the first λ_{start} for the fused lasso is different in each fold and when the maximum number of steps is reached before the path reached λ_{end} the last evaluated (or found) λ_{last}. Is not guaranteed to be the same across fold. So for the fused lasso we can not define the regularization values to evaluate, but must rather inspect the results for each fold. We solve this here rather pragmatically. We inspect MSE lines across the solution path for each fold. For the overlapping region we define the mean of all evaluated points and choose $\lambda_m in$ as the amount of regularization that minimizes the MSE for common area of the solution path. The λ values will not exactly be the same for the folds, so we have to interpolate the gaps between the actual λ, to create a common range we can compute the mean on. The interpolation is done via local polynomial regression fitting using the *loess* function in R with a span of 0.05 and degree 2 (Cleveland, Grosse, and Shyu (1992)).

6.4 Results

6.4.1 Fused lasso



The error lines in the different folds differ in their trajectories as well as in their starting points (**??**). Note that we cut off fold 3 for better readability of the plot (the MSE reaches until 3000). The black line indicates the mean, computed for the area that is covered by all error lines (after interpolation).

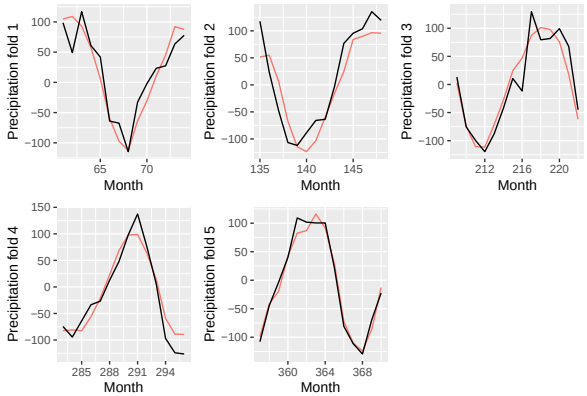


Figure 6.1: Precipitation prediction and target values in the test set in each fold. Predictions in red and target values in black.

The predictions inside the folds are very similar to lasso without standardization (see **??**), the same holds for the predictions from the full model, but the MSE improves here (@**??**fig:pred-plot-full-fused-og).

```
## [1] 1131.709
```

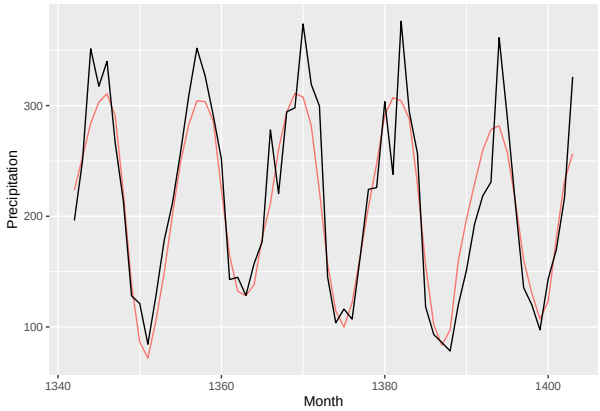
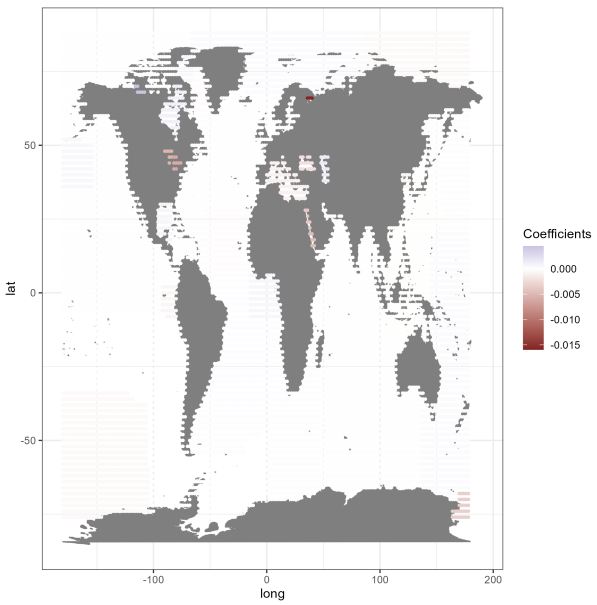
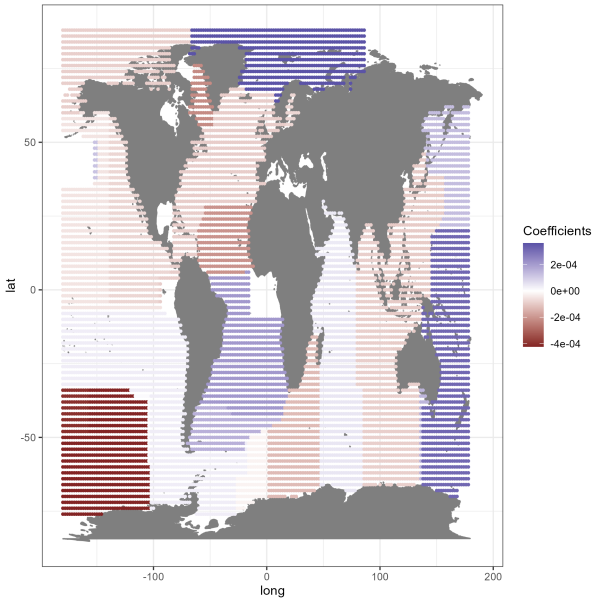


Figure 6.2: Precipitation prediction and target values in the validation set. Predictions in red and target values in black. The model was fitted on the full CV data with the lambda value that minimised the average MSE





Cleveland, WS, E Grosse, and WM Shyu. 1992. "Local Regression Models. Chapter 8 in Statistical Models in s (JM Chambers and TJ Hastie Eds.), 608 p." *Wadsworth & Brooks/Cole, Pacific Grove, CA*.