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

3.1 Short Recap

We give a short overview over the correlation between monthly sea surface temperature and monthly mean precipitation in the Central Amazonas Basin (CAB). First we will analyse the original and then the deaseasonalised data. SST and precipitation data have been deseasonalised, meaning first each time series was decomposed by the stl algorithm according to

$$Monthly\ Data = Seasonal + Trend + Remainder$$

Afterwards only trends and remainders time series were kept to constitute a new time series that will be used as predictor (SST) and target (precipitation).

In a next step we compute the pearson correlation coefficient ρ between each SST grid point time series and the mean precipitation time series.

$$\rho_{X,Y} = \frac{\mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right]}{\sigma_X \sigma_Y} \tag{3.1}$$

Since our goal is to predict the precipitation on the SST information, we are also interested in the correlation of the SST and future precipitation some months ahead. To examine this we also compute the correlations for different time lags. For example for a time lag of 6 month we correlate January SST data and precipitation in July.

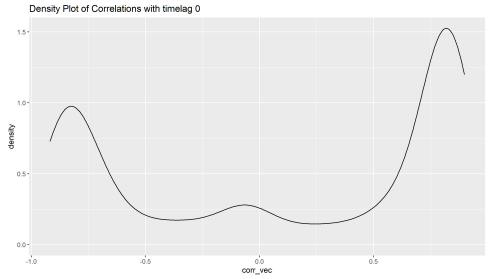
We consider time lags of 0,3,6 and 12 months. And show the density of the correlation values as well as their spatial distribution on a map. We also display the highest positive and negative correlation based on their respective 2.5% and 97.5% quantiles. All correlations that are between these values are set to 0 then.

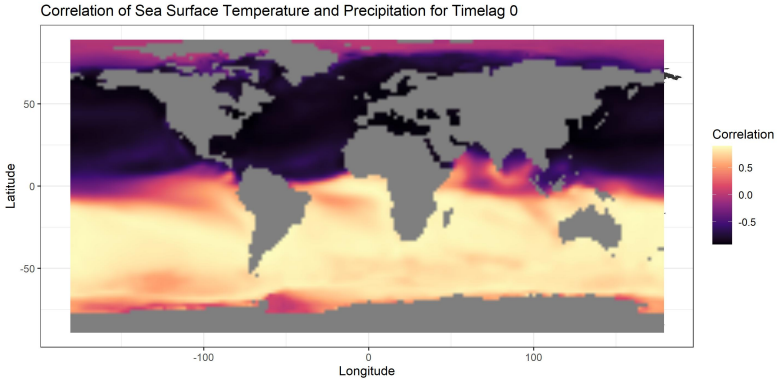
3.2 Correlation of Sea Surface Temperature and Precipitation

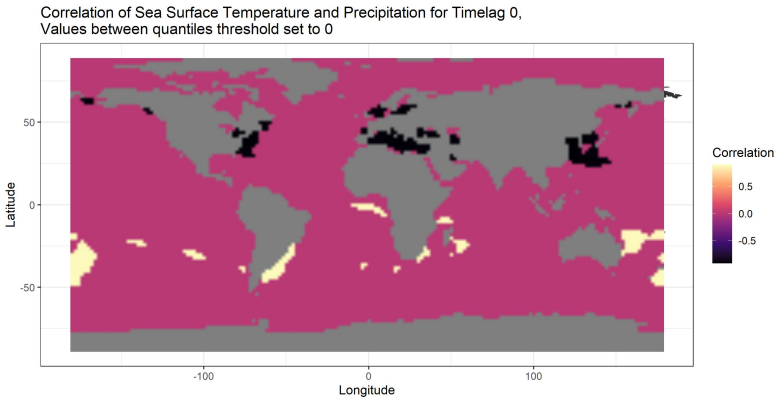
3.2.1 Original Data

Following, for each timelag we show the respective density of correlation values, their location on the map and also the 5% strongest positive and negative correlations.

3.2.1.1 Timelag 0

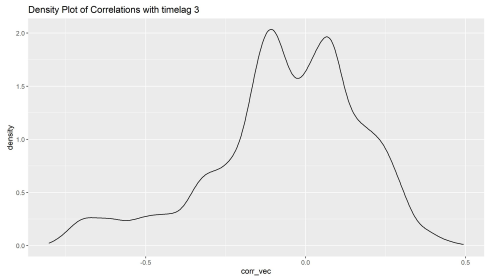


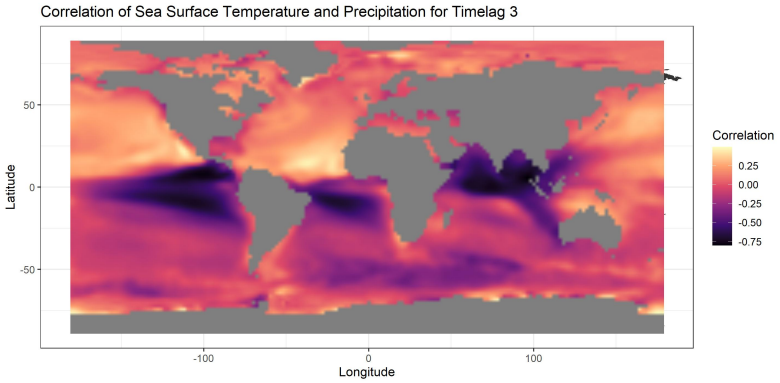


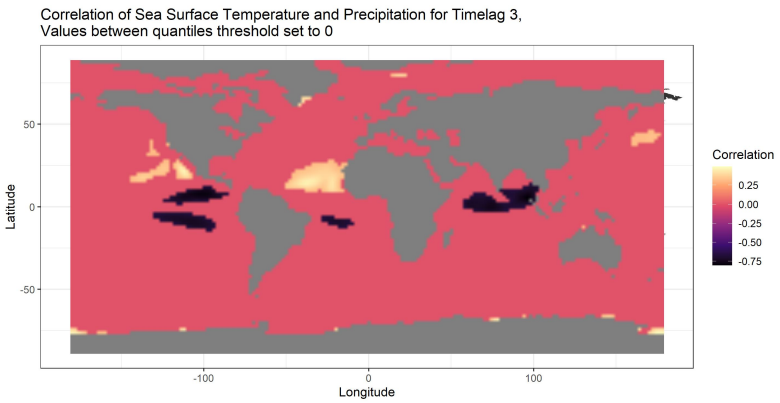


Inspecting the density plot for time lag 0, we see two modi for correlations, one for negative correlations around -0.8 and one for positive correlations around 0.8. Also a small spike can be seen for low negative correlations. If we plot these correlations on the respective grid points we see a clear north-south negative-positive correlation distinction. The “boarder” is organised around the equator. The plot for the strongest 5% of correlations reveals areas with strong positive and negative correlations in the north and south respectively.

3.2.1.2 Timelag 3

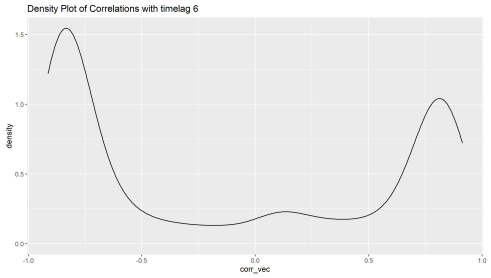


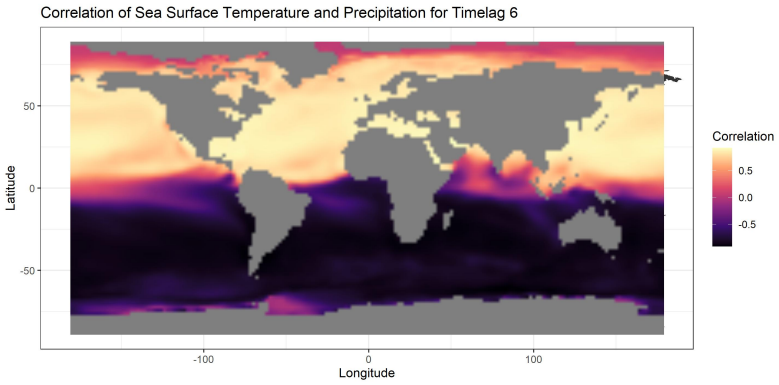


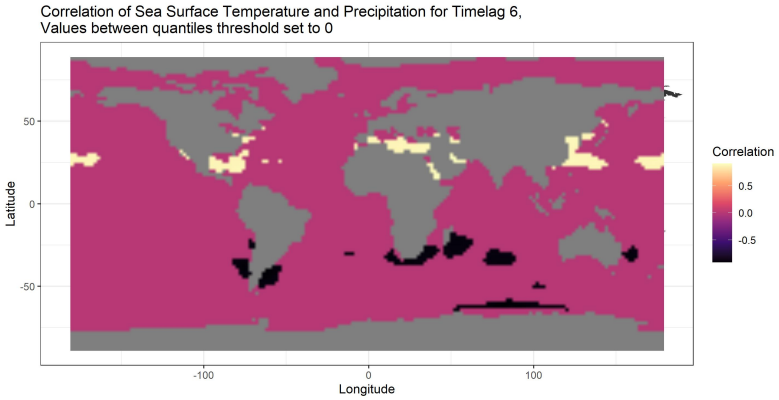


The density of correlations for timelag 3, is left-skewed and has two modi that are organised around 0 and -0.125 respectively. The correlation map shows that the high positive and negative correlations are more close to equator here. Note that the legend for the correlationmap is “shifted” here, because the maximal negative correlation has a higher absolute value than the maximal positive correlation. The strongest correlations also seem to be shifted towards the equator.

3.2.1.3 Timelag 6

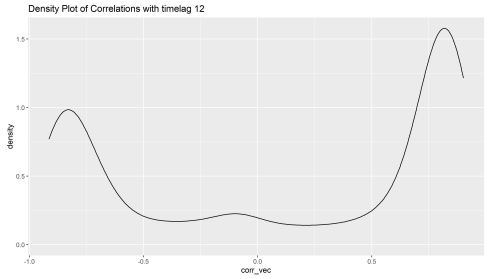


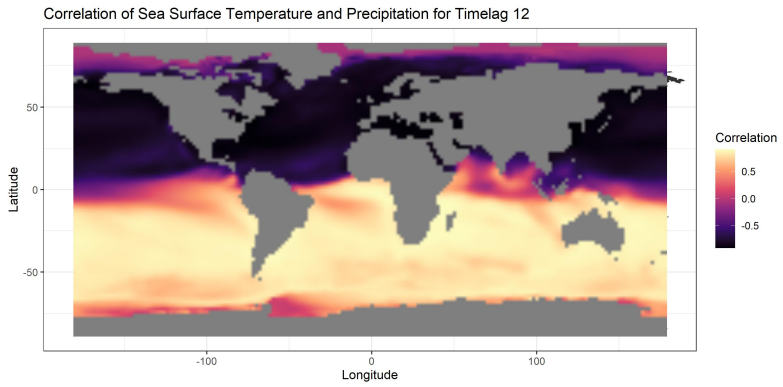


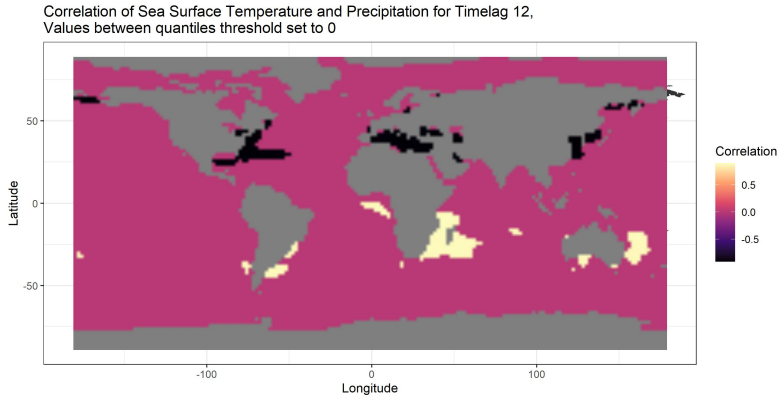


We can see the density plot for timelag 6 is pretty similar to the one of timelag 0 but seems to be "flipped" around 0. Similarly the correlation map shows (high) negative correlations in the south now and high postive correlations in the north.

3.2.1.4 Timelag 12





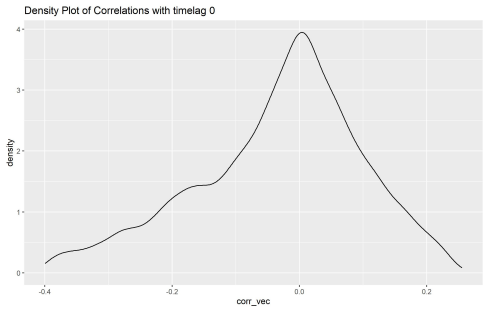


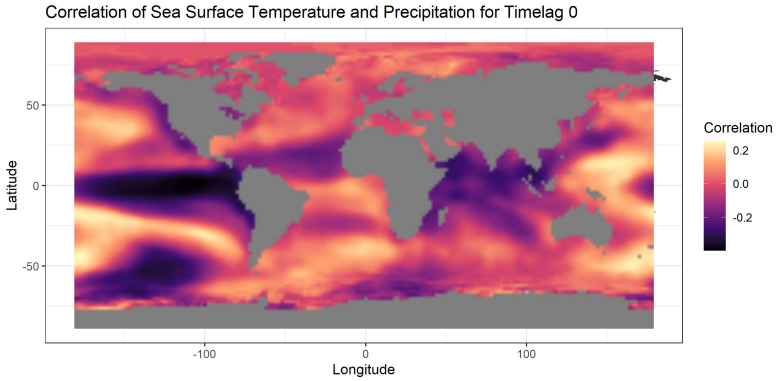
Giving a time lag of one year, we can see that the distribution of correlations is now again similar to the distribution for time lag 0. This also hold for the location of positive and negative correlations in general, as well as for the strongest 5% of correlations.

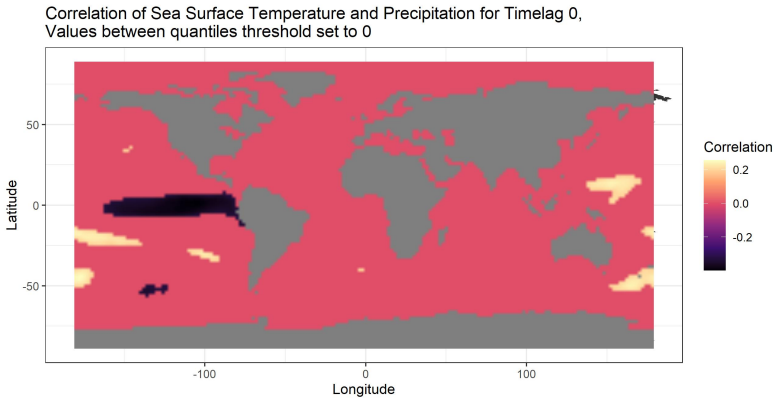
3.2.2 Deseasonalised Data

Following, for each time lag we show the respective density of correlation values, their location on the map and also the 5% strongest positive and negative correlations.

3.2.2.1 Timelag 0

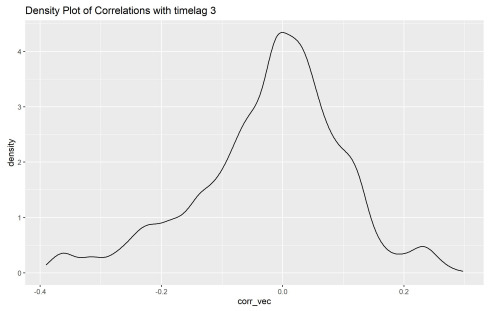


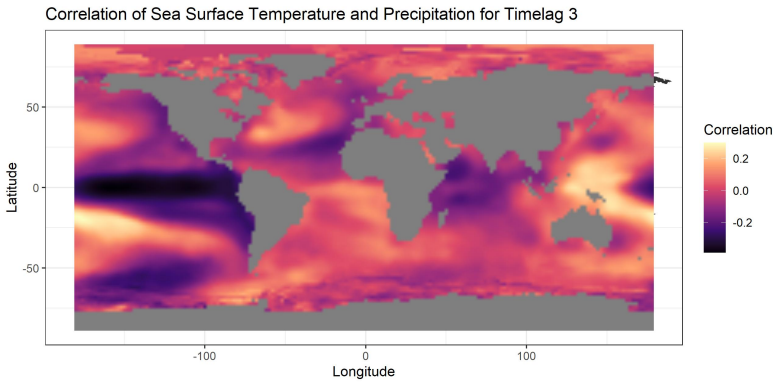


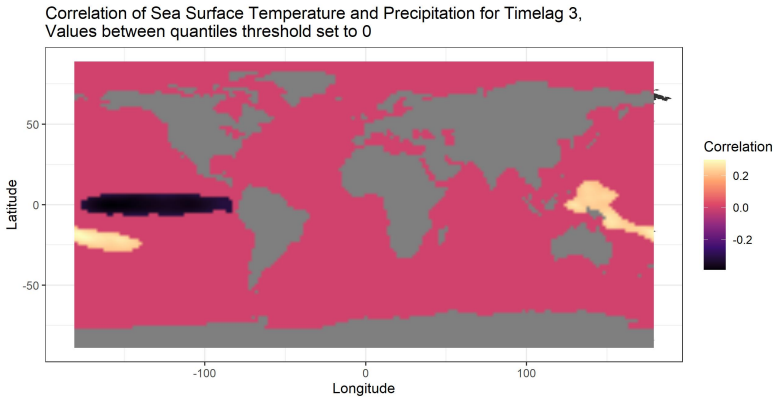


Inspecting the density plot for timelag 0, we see that after excluding seasonality from the time series we get a left-skewed distribution of correlations. With a mode around 0. In general the correlation values are a lot lower than in the original data. With a maximum at around -0.4 and +2.5 respectively. We plot these correlations on the respective grid and see that the clear north south distinction in the correlations before deseasonalising the data does not appear anymore. The plot for the strongest 5% of correlations reveals areas with strongest positive and negative correlations. But as stated before the values are in general much lower.

3.2.2.2 Timelag 3

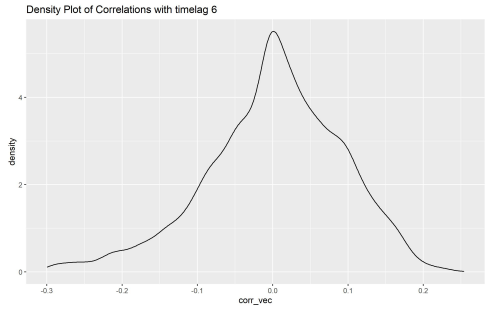


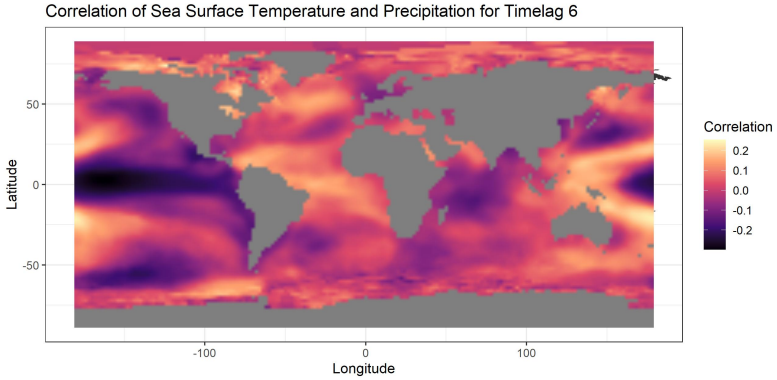


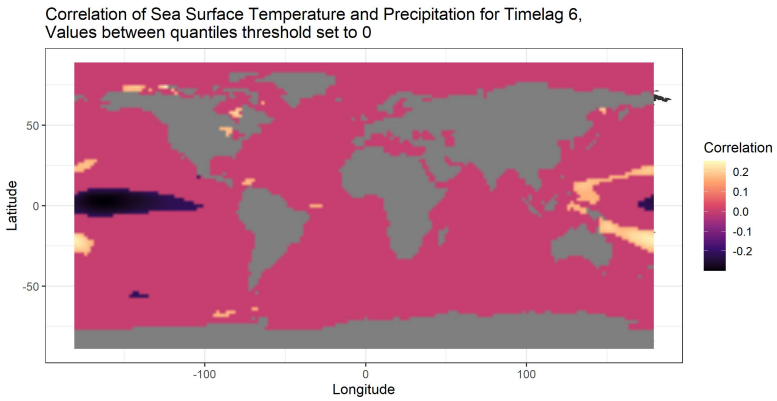


For the density of timelag 3 we get a similar picture as for timelag 0. The mode is a higher and the tails get a bit more mass. Also the correlation map does not seem to change a lot. The strongest correlations appear to be shifted to the left.

3.2.2.3 Timelag 6

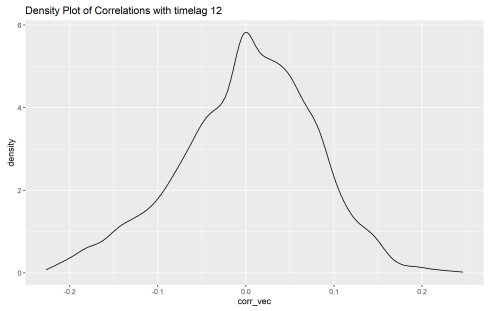


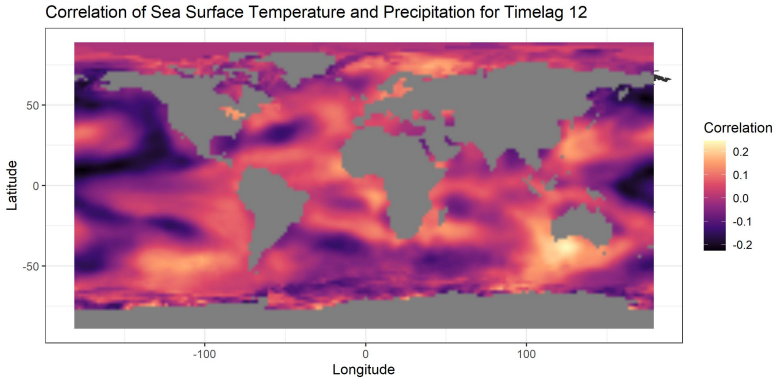


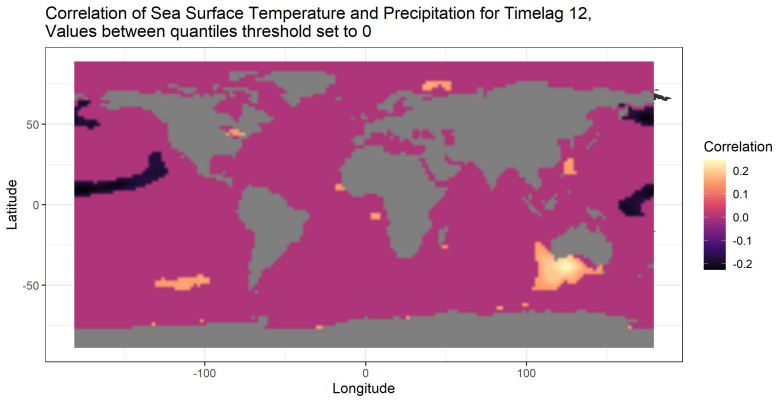


Since the distributions of correlation values are all unimodal we do not observe the "flip" we saw in the original data when comparing the densities of timelag 0 and 6. The mode again gets larger and the maximum postive and negative correlation values get smaller. The strongest negative correlations are shifted further to left.

3.2.2.4 Timelag 12







Given a timelag of one year, the distribution now has a mode around 6, and started at around 4 when timelag was 0. Also neither the positive nor negative correlations exceed values of 2.5. The consistent region of strong negative and positive correlations is now less organised or more scattered.

3.3 Summary

3.3.1 Original Data

We can observe that the positive and negative correlations of sst and precipitation follow a spatial and temporal pattern. The location and density of the positive and negative correlation "wanders" over the equator in opposite directions. The densities and correlationmaps for timelag 0 and 6 appear to be quite similar but "flipped". The densities and correlationmaps for timelag 0 and 12 appear again to be similar. The same pattern seems to hold for the strongest correlations.

3.3.2 Deseasonalised Data

Correlation values are in general a lot lower than in the original data and decrease with increasing timelag. We still observe temporal and regional patterns, although these dissolve a bit for a timelag of 12.

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