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

Dario Lepke

September 09, 2022

1. Introduction
2. Explorative analysis
3. Correlation analysis
4. Clustering
5. The lasso
6. lasso evaluation
7. lasso settings

8. lasso results
9. The fused lasso
10. fused evaluation
11. fused settings
12. fused results
13. Summary

Introduction

- The Amazon basin is a key hotspot of biodiversity, carbon storage and moisture recycling
- Hydrological extremes affect ecosystem and populations tremendously
- Droughts in the Amazon rainforest can have severe biomass carbon impact
- Severe Amazon drought in 2010 had total biomass carbon impact of 2.2 PgC, affected area $3 miokm^2$

- Ciemer et al. (2020) established an early warning indicator for water deficits in the central Amazon basin (CAB)

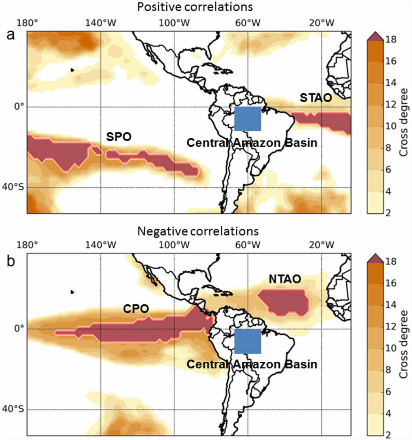


Figure 1: Cross degree between sea surface temperature and continental rainfall anomalies. For each grid cell of sea surface temperature in the Atlantic and Pacific, the cross degree towards rainfall in the Central

- Inspect spatial and temporal characteristics in raw data
- Directly predict rain from SST
- Use lasso and fused lasso
- model evaluation with cross validation for time series

Explorative analysis

- Rain data from CHIRPS ()
- CHIRPS contains in-situ and satellite data
- SST data from ERSST (Extended Reconstructed Sea Surface Temperature)
- ERSST is reanalysis of observation data (made by ships and buoys for example), missing data filled by interpolation techniques
- These are the same data sets as in Ciemer et al. (2020)

- show area
- show mean and sd
- show glyph plots

- show mean and sd

Correlation analysis

- show timelag 0, raw and de-seasonalized

Clustering

- explorative analysis has shown spatial and temporal differences in the precipitation data
- we explored this further using k-means clustering
- steps: find optimal k via pca and gap statistic
- apply k-means to original precipitation data
- we compared k-means and k-medoid with and without PCA via the gap statistic
- here show only k-means with PCA as it gave best results
- applying the regression models to separate clusters might improve predictions
- Using 3 principal components and 5 cluster centers with k-means gave best results on gap statistic

- Our objective is to find k internally homogeneous and externally heterogeneous clusters
- Similarity is measured by the euclidean distance

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = ||x_i - x_{i'}||^2 \tag{1}$$

- And we want to minimize the sum of distances inside all clusters, given by:

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} ||x_i - x_{i'}||^2 = \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - \bar{x}_k||^2 \tag{2}$$

where $\bar{x} = (\bar{x}_{1k}, ..., \bar{x}_{pk})$ stands for the mean vectors of the kth cluster and $N_k = \sum_{i=1}^{N} I(C(i) = k)$.

- number of clusters has to be defined beforehand
- we decided on the optimal number of k using the gap statistic
- Let W_k be $W(C)$ for fix k
- We compare W_k from the precipitation data with average W_k^* from B Monte Carlo sampled data sets

$$Gap(k) = E\{log(W*_k)\} - log(W_k). \tag{3}$$

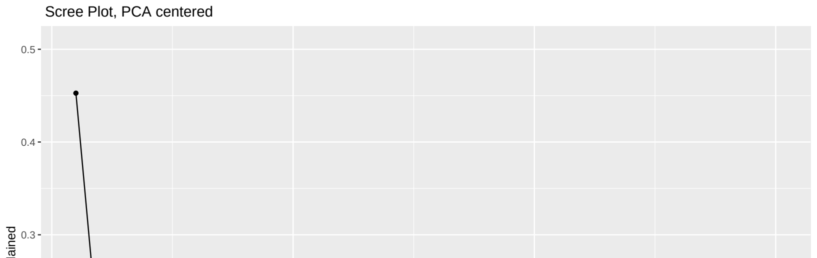
- We choose k as smallest k such that

$$Gap(k) \geq Gap(k+1) - s_{k+1} \tag{4}$$

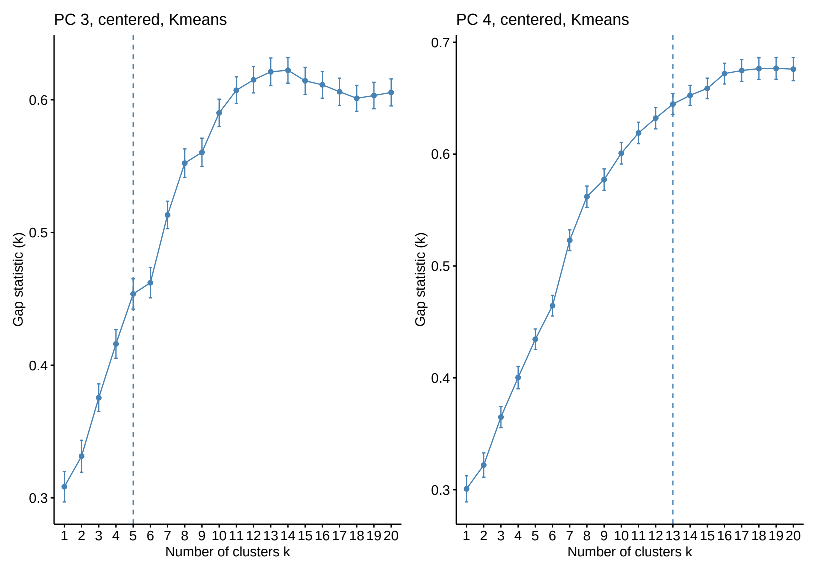
- s_{k+1} is $sd_k\sqrt{1+1/B}$, and sd the standard deviation of log(W*_k)

- Before running k-means we center the precipitation data and apply a PCA to reduce the large number of correlated variables to a few
- The new variables are linear combinations of the original variables
- Here: Each variable is a month of precipitation data in the CAB

```
## Scale for 'y' is already present. Adding another scale f
## replace the existing scale.
```



- The “elbow” be observe in the screeplot suggest 3 or 4 principal components
- The first 3 and 4 first PC explain 67.77 and 70.79 of the variance respectively.
- We compare the gap statistic results for 3 and 4 PC



- The k-means gap statistic on the first 3 PC proposes 5 clusters - For 4 PC, 13 clusters are chosen

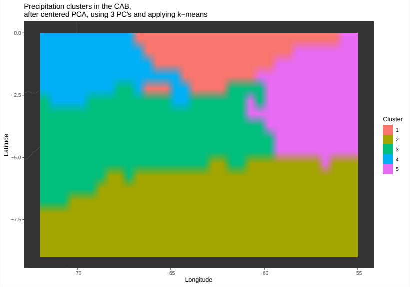

Precipitation clusters in the CAB,
after centered PCA, using 3 PC's and applying k−means
0.0
−2.5
−5.0
−7.5
Latitude
−70
−65
−60
−55
Longitude
Cluster
1
2
3
4
5

Figure 3: Spatial distribution of the found clusters in the CAB. We applied a centered PCA on the data and used 3 principal components before applying the k-means algorithm

- We find 5 clusters of different sizes
- The found clusters are almost completely spatially coherent although we did not include any spatial dependencies in the clustering
- Small exception is the “island” of cluster 1 (orange) inside

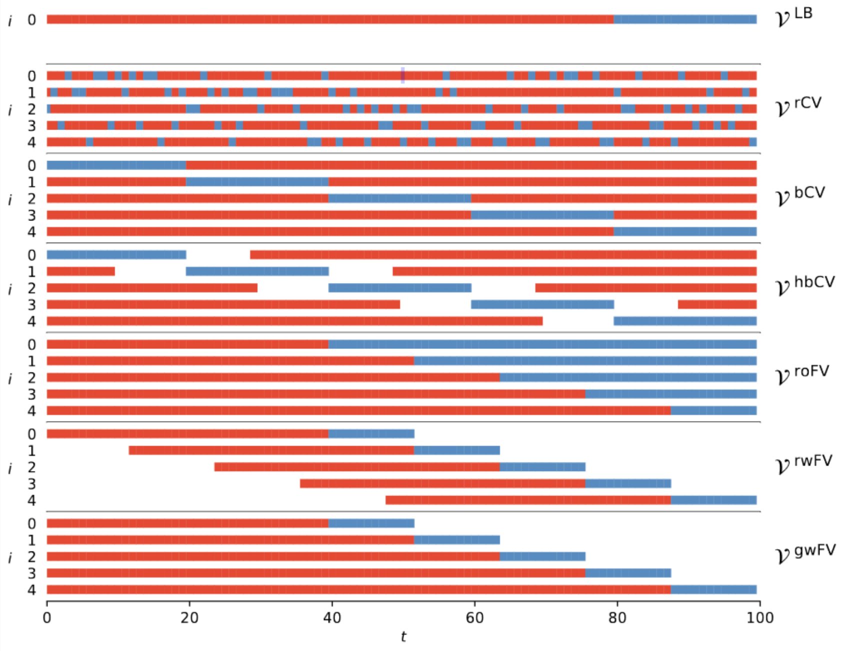
The lasso

- We now consider the lasso regression problem

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1] \quad (5)$$

- In our setting n « p, so lasso is natural choice
- The problem is solved using coordinate descent
- Due to the time dependencies in our data normal Cross Validation may be unjustified

- Our goal is to train a model that can also predict well on new, unseen data
- We simulate the situation of unseen data by splitting our data into one part for model selection and another part for model evaluation
- Model evaluation is usually done via Cross Validation, but classic Cross Validation does not take into account the time dependency in our data



i
0
$\mathcal{V}^{\text{LB}}$
0
1
2
3
4
$\mathcal{V}^{\text{rCV}}$
$\mathcal{V}^{\text{bCV}}$
$\mathcal{V}^{\text{hbCV}}$
$\mathcal{V}^{\text{roFV}}$
$\mathcal{V}^{\text{rwFV}}$
$\mathcal{V}^{\text{gwFV}}$
0
20
40
60
80
100
t

lasso evaluation

lasso settings

lasso results

The fused lasso

- why fused
- what is fused
- optimization

fused evaluation

fused settings

fused results

Summary

Important test

Ciemer, Catrin, Lars Rehm, Juergen Kurths, Reik V Donner, Ricarda Winkelmann, and Niklas Boers. 2020. “An Early-Warning Indicator for Amazon Droughts Exclusively Based on Tropical Atlantic Sea Surface Temperatures.” *Environmental Research Letters* 15 (9): 094087.