

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

Dario Lepke

Contents

Introduction	13
1 Related work	15
2 EDA precipitation	17
2.1 Overview	17
2.2 Precipitation values raw	17
2.3 Mean at each location	17
2.4 SD at each location	17
2.5 Mean and SD at each location	17
2.6 Trend at each location	17
2.7 Means per month TEST	17
2.8 SD per month TEST	17
2.9 Trend per month TEST	17
3 EDA SST	19
3.1 SST values raw	19
3.2 Mean at each location	19
3.3 SD at each location	19
3.4 Mean and SD at each location	19
3.5 Trend at each location	19
3.6 Means per month TEST	19
3.7 SD per month TEST	19
3.8 Trend per month TEST	19

4	Glyph plots sst	21
4.1	For smaller window	21
5	Glyph plots	23
6	Correlation analysis	25
6.1	Short Recap	25
6.2	Correlation of Sea Surface Temperature and Precipitation	25
6.3	Summary	26
7	Main Idea Clustering	27
8	Clustering Methods	29
8.1	K-means	29
8.2	K-medoids	29
8.3	PCA	29
8.4	Gap statistic	29
9	Clustering results	31
9.1	K-means and PAM gap statistics without PCA	31
9.2	K-means and PAM gap statistics after applying PCA	31
9.3	Summary	31
10	Analyse clustering results	33
11	Fused on OG	35
11.1	Error plots	35
11.2	Coefficient plots	35
11.3	Inspect predictions from each fold	35
12	Lasso on original data	37
12.1	LASSO model	37
12.2	Error plots	37
12.3	Coefficient plots	37

<i>CONTENTS</i>	5
12.4 Inspect predictions from each fold	37
12.5 Inspect predictions from best CV-lambda	37
12.6 Summary	37
13 Lasso on cluster 1	39
13.1 Cluster 1	39
13.2 Error plots	39
13.3 Coefficient plots	39
13.4 Inspect predictions from each fold	39
13.5 Inspect predictions from best CV-lambda	39
13.6 Summary	39
14 Lasso with standardization	41
14.1 Model	41
14.2 Error plots	41
14.3 Coefficient plots	41
14.4 Inspect predictions from each fold	41
14.5 Inspect predictions from best CV-lambda	41
14.6 Summary	41
15 Lasso with standardization updated	43
15.1 Model	43
15.2 Error plots	43
15.3 Coefficient plots	43
15.4 Inspect predictions from each fold	43
15.5 Inspect predictions from best CV-lambda	43
15.6 Summary	43
16 noclust lasso	45
16.1 LASSO model	45
16.2 Error plots	45
16.3 Coefficient plots	45
16.4 Inspect predictions from each fold	45

16.5	Inspect predictions from best CV-lambda	45
16.6	Summary	45
17	noclust lasso stand	47
17.1	LASSO model	47
17.2	Error plots	47
17.3	Coefficient plots	47
17.4	Inspect predictions from each fold	47
17.5	Inspect predictions from best CV-lambda	47
17.6	Summary	47
18	Lasso with center	49
18.1	Model	49
18.2	Error plots	49
18.3	Coefficient plots	49
18.4	Inspect predictions from each fold	49
18.5	Inspect predictions from best CV-lambda	49
18.6	Summary	49
19	Lasso on cluster 2	51
19.1	Cluster 2	51
19.2	Error plots	51
19.3	Coefficient plots	51
19.4	Inspect predictions from each fold	51
19.5	Inspect predictions from best CV-lambda	51
19.6	Summary	51
20	Lasso on cluster 3	53
20.1	Cluster 3	53
20.2	Error plots	53
20.3	Coefficient plots	53
20.4	Inspect predictions from each fold	53
20.5	Inspect predictions from best CV-lambda	53
20.6	Summary	53

<i>CONTENTS</i>	7
21 Lasso on cluster 4	55
21.1 Cluster 4	55
21.2 Error plots	55
21.3 Coefficient plots	55
21.4 Inspect predictions from each fold	55
21.5 Inspect predictions from best CV-lambda	55
21.6 Summary	55
22 Lasso on cluster 5	57
22.1 Cluster 5	57
22.2 Error plots	57
22.3 Coefficient plots	57
22.4 Inspect predictions from each fold	57
22.5 Inspect predictions from best CV-lambda	57
22.6 Summary	57
23 lasso with lags	59
23.1 LASSO model with lags	59
23.2 Error plots	59
23.3 Coefficient plots	59
23.4 Inspect predictions from each fold	59
23.5 Inspect predictions from best CV-lambda	59
23.6 Summary	59
24 lasso with diff	61
24.1 LASSO model with diff	61
24.2 Error plots	61
24.3 Coefficient plots	61
24.4 Inspect predictions from each fold	61
24.5 Inspect predictions from best CV-lambda	61
24.6 Summary	61

25 lasso with diff 2	63
25.1 LASSO model with diff	63
25.2 Error plots	63
25.3 Coefficient plots	63
25.4 Inspect predictions from each fold	63
25.5 Inspect predictions from best CV-lambda	63
25.6 Summary	63
26 lasso with deseasonalised	65
26.1 LASSO model, stl() on SST	65
26.2 Error plots	65
26.3 Coefficient plots	65
26.4 Inspect predictions from each fold	65
26.5 Inspect predictions from best CV-lambda	65
26.6 Summary	65
27 Lasso on sst anomalies	67
27.1 LASSO model	67
27.2 Error plots	67
27.3 Coefficient plots	67
27.4 Inspect predictions from each fold	67
27.5 Inspect predictions from best CV-lambda	67
27.6 Summary	67
28 Lasso on sst anomalies, standardized	69
28.1 LASSO model	69
28.2 Error plots	69
28.3 Coefficient plots	69
28.4 Inspect predictions from each fold	69
28.5 Inspect predictions from best CV-lambda	69
28.6 Summary	69

29 Lasso on sst anomalies, differentiated	71
29.1 LASSO model	71
29.2 Error plots	71
29.3 Coefficient plots	71
29.4 Inspect predictions from each fold	71
29.5 Inspect predictions from best CV-lambda	71
29.6 Summary	71
30 Lasso on sst anomalies, differentiated and standardized	73
30.1 LASSO model	73
30.2 Error plots	73
30.3 Coefficient plots	73
30.4 Inspect predictions from each fold	73
30.5 Inspect predictions from best CV-lambda	73
30.6 Summary	73
31 Small fused on OG	75
31.1 Error plots	75
31.2 Coefficient plots	75
31.3 Inspect predictions from each fold	75
32 Small fused stand	77
32.1 Error plots	77
32.2 Coefficient plots	77
32.3 Inspect predictions from each fold	77
33 Fused on OG gamma 0.1	79
33.1 Error plots	79
33.2 Coefficient plots	79
33.3 Inspect predictions from each fold	79

34 Fused on OG gamma 0.1 stand	81
34.1 Error plots	81
34.2 Coefficient plots	81
34.3 Inspect predictions from each fold	81
35 explain gamma 01 results	83
36 Fused 5k	85
36.1 Error plots	85
36.2 Coefficient plots	85
36.3 Inspect predictions from each fold	85
37 Noclust Fused 5k	87
37.1 Error plots	87
37.2 Coefficient plots	87
37.3 Inspect predictions from each fold	87
38 Fused 5k stand	89
38.1 Error plots	89
38.2 Coefficient plots	89
38.3 Inspect predictions from each fold	89
39 Fused on OG gamma 0.1 stand	91
39.1 Error plots	91
39.2 Coefficient plots	91
39.3 Inspect predictions from each fold	91
40 LASSO Regression	93
40.1 The LASSO	93
40.2 TODO here	93
41 Fused Lasso Regression	95
41.1 The fused lasso	95
42 quick summary what I got so far from the algorithm	97

<i>CONTENTS</i>	11
43 what to they actually use now	99
44 test lasso new plots	101
44.1 LASSO model	101
44.2 Error plots	101
44.3 Coefficient plots	101
44.4 Inspect predictions from each fold	101
44.5 Inspect predictions from best CV-lambda	101
44.6 Summary	101

Introduction

Placeholder

Chapter 1

Related work

Placeholder

Chapter 2

EDA precipitation

Placeholder

2.1 Overview

2.2 Precipitation values raw

2.3 Mean at each location

2.4 SD at each location

2.5 Mean and SD at each location

2.6 Trend at each location

2.7 Means per month TEST

2.8 SD per month TEST

2.9 Trend per month TEST

Chapter 3

EDA SST

Placeholder

3.1 SST values raw

3.2 Mean at each location

3.3 SD at each location

3.4 Mean and SD at each location

3.5 Trend at each location

3.6 Means per month TEST

3.7 SD per month TEST

3.8 Trend per month TEST

Chapter 4

Glyph plots sst

Placeholder

4.1 For smaller window

Chapter 5

Glyph plots

Placeholder

Chapter 6

Correlation analysis

Placeholder

6.1 Short Recap

6.2 Correlation of Sea Surface Temperature and Precipitation

6.2.1 Original Data

6.2.1.1 Timelag 0

6.2.1.2 Timelag 3

6.2.1.3 Timelag 6

6.2.1.4 Timelag 12

6.2.2 Deseasonalised Data

6.2.2.1 Timelag 0

6.2.2.2 Timelag 3

6.2.2.3 Timelag 6

6.2.2.4 Timelag 12

6.3 Summary

6.3.1 Original Data

6.3.2 Deseasonalised Data

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.

Chapter 7

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 $5^\circ \times 5^\circ$), we have 612 cells and 432 months, resulting in a 612×432 data matrix. we want to cluster cells into homogenous groups.

Chapter 8

Clustering Methods

Placeholder

8.1 K-means

8.1.1 Kmeans characteristics

8.2 K-medoids

8.2.1 K-medoids characteristics

8.3 PCA

8.4 Gap statistic

Chapter 9

Clustering results

Placeholder

9.1 K-means and PAM gap statistics without PCA

9.1.1 Scree plot

9.2 K-means and PAM gap statistics after applying PCA

9.3 Summary

Chapter 10

Analyse clustering results

Placeholder

Chapter 11

Fused on OG

Placeholder

11.1 Error plots

11.2 Coefficient plots

11.3 Inspect predictions from each fold

Chapter 12

Lasso on original data

Placeholder

12.1 LASSO model

12.2 Error plots

12.3 Coefficient plots

12.4 Inspect predictions from each fold

12.5 Inspect predictions from best CV-lambda

12.6 Summary

Chapter 13

Lasso on cluster 1

Placeholder

13.1 Cluster 1

13.2 Error plots

13.3 Coefficient plots

13.4 Inspect predictions from each fold

13.5 Inspect predictions from best CV-lambda

13.6 Summary

Chapter 14

Lasso with standardization

Placeholder

14.1 Model

14.2 Error plots

14.3 Coefficient plots

14.4 Inspect predictions from each fold

14.5 Inspect predictions from best CV-lambda

14.6 Summary

Chapter 15

Lasso with standardization updated

Placeholder

15.1 Model

15.2 Error plots

15.3 Coefficient plots

15.4 Inspect predictions from each fold

15.5 Inspect predictions from best CV-lambda

15.6 Summary

Chapter 16

noclust lasso

Placeholder

16.1 LASSO model

16.2 Error plots

16.3 Coefficient plots

16.4 Inspect predictions from each fold

16.5 Inspect predictions from best CV-lambda

16.6 Summary

Chapter 17

noclust lasso stand

Placeholder

17.1 LASSO model

17.2 Error plots

17.3 Coefficient plots

17.4 Inspect predictions from each fold

17.5 Inspect predictions from best CV-lambda

17.6 Summary

Chapter 18

Lasso with center

Placeholder

18.1 Model

18.2 Error plots

18.3 Coefficient plots

18.4 Inspect predictions from each fold

18.5 Inspect predictions from best CV-lambda

18.6 Summary

Chapter 19

Lasso on cluster 2

Placeholder

19.1 Cluster 2

19.2 Error plots

19.3 Coefficient plots

19.4 Inspect predictions from each fold

19.5 Inspect predictions from best CV-lambda

19.6 Summary

Chapter 20

Lasso on cluster 3

Placeholder

20.1 Cluster 3

20.2 Error plots

20.3 Coefficient plots

20.4 Inspect predictions from each fold

20.5 Inspect predictions from best CV-lambda

20.6 Summary

Chapter 21

Lasso on cluster 4

Placeholder

21.1 Cluster 4

21.2 Error plots

21.3 Coefficient plots

21.4 Inspect predictions from each fold

21.5 Inspect predictions from best CV-lambda

21.6 Summary

Chapter 22

Lasso on cluster 5

Placeholder

22.1 Cluster 5

22.2 Error plots

22.3 Coefficient plots

22.4 Inspect predictions from each fold

22.5 Inspect predictions from best CV-lambda

22.6 Summary

Chapter 23

lasso with lags

Placeholder

23.1 LASSO model with lags

23.2 Error plots

23.3 Coefficient plots

23.4 Inspect predictions from each fold

23.5 Inspect predictions from best CV-lambda

23.6 Summary

Chapter 24

lasso with diff

Placeholder

24.1 LASSO model with diff

24.2 Error plots

24.3 Coefficient plots

24.4 Inspect predictions from each fold

24.5 Inspect predictions from best CV-lambda

24.6 Summary

Chapter 25

lasso with diff 2

Placeholder

25.1 LASSO model with diff

25.2 Error plots

25.3 Coefficient plots

25.4 Inspect predictions from each fold

25.5 Inspect predictions from best CV-lambda

25.6 Summary

Chapter 26

lasso with deseasonalised

Placeholder

26.1 LASSO model, `stl()` on SST

26.2 Error plots

26.3 Coefficient plots

26.4 Inspect predictions from each fold

26.5 Inspect predictions from best CV-lambda

26.6 Summary

Chapter 27

Lasso on sst anomalies

Placeholder

27.1 LASSO model

27.2 Error plots

27.3 Coefficient plots

27.4 Inspect predictions from each fold

27.5 Inspect predictions from best CV-lambda

27.6 Summary

Chapter 28

Lasso on sst anomalies, standardized

Placeholder

28.1 LASSO model

28.2 Error plots

28.3 Coefficient plots

28.4 Inspect predictions from each fold

28.5 Inspect predictions from best CV-lambda

28.6 Summary

Chapter 29

Lasso on sst anomalies, differentiated

Placeholder

29.1 LASSO model

29.2 Error plots

29.3 Coefficient plots

29.4 Inspect predictions from each fold

29.5 Inspect predictions from best CV-lambda

29.6 Summary

Chapter 30

Lasso on sst anomalies, differentiated and standardized

Placeholder

30.1 LASSO model

30.2 Error plots

30.3 Coefficient plots

30.4 Inspect predictions from each fold

30.5 Inspect predictions from best CV-lambda

30.6 Summary

Chapter 31

Small fused on OG

Placeholder

31.1 Error plots

31.2 Coefficient plots

31.3 Inspect predictions from each fold

Chapter 32

Small fused stand

Placeholder

32.1 Error plots

32.2 Coefficient plots

32.3 Inspect predictions from each fold

Chapter 33

Fused on OG gamma 0.1

Placeholder

33.1 Error plots

33.2 Coefficient plots

33.3 Inspect predictions from each fold

Chapter 34

Fused on OG gamma 0.1 stand

Placeholder

34.1 Error plots

34.2 Coefficient plots

34.3 Inspect predictions from each fold

Chapter 35

explain gamma 01 results

Placeholder

Chapter 36

Fused 5k

Placeholder

36.1 Error plots

36.2 Coefficient plots

36.3 Inspect predictions from each fold

Chapter 37

Noclust Fused 5k

Placeholder

37.1 Error plots

37.2 Coefficient plots

37.3 Inspect predictions from each fold

Chapter 38

Fused 5k stand

Placeholder

38.1 Error plots

38.2 Coefficient plots

38.3 Inspect predictions from each fold

Chapter 39

Fused on OG gamma 0.1 stand

Placeholder

39.1 Error plots

39.2 Coefficient plots

39.3 Inspect predictions from each fold

Chapter 40

LASSO Regression

Placeholder

40.1 The LASSO

40.2 TODO here

Chapter 41

Fused Lasso Regression

41.1 The fused lasso

```
# TODO add statistical learning with sparsity to references.
```

```
# TODO add to summary of LASSO defaults, that characteristic  
# of LASSO
```

As expected and seen in the results, the different LASSO models choose single SST regions as predictors as opposed to whole regions. Since the LASSO only regularizes the magnitude of coefficients but ignores their ordering. We therefore use the so-called *fused lasso* as implemented in the *genlasso* package and the respective *fusedlasso* function ([@ref{genlassopackage}](#), [@ref{thibshirani2005sparsity}](#)).

```
# TODO make reference to achieved results here
```

```
# TODO cite general LASSO properties here  
# how it acts in situations where coefficients are grouped  
# or highly correlated.
```

```
# TODO add plot of the graph given by our helper functions.
```

```
# TODO add graphic of fig.2 fused lasso
```

The fused LASSO solves the problem:

$$\min_{\beta} 1/2 \sum_{i=1}^n (y_i - x_i^T \beta_i)^2 + \lambda \sum_{i,j \in E} |\beta_i - \beta_j| + \gamma \cdot \lambda \sum_{i=1}^p |\beta_i|,$$

with x_i being the i th row of the predictor matrix and E is the edge set of an underlying graph. Regularizing $|\beta_i - \beta_j|$, penalizes large differences in close coefficients. In our case “close” means small distances as defined on 2-dimensional longitude/latitude grid. This grid defines a graph that can be used to compute the differences for each location. The third term $\gamma \cdot \lambda \sum_{i=1}^p |\beta_i|$, controls the sparsity of the coefficients. $\gamma = 0$ leads to complete fusion of the coefficients (no sparsity) and $\gamma > 0$ introduces sparsity to the solution, with higher values placing more priority on sparsity. $\hat{\beta}$ is computed as a function of λ , with fixed γ .

#TODO note the fused lasso paper here are

*# TODO add efficient implementations of hte generlaized lasso dual path
algorithm to references*

TODO ?check generalized dual path algorithm og paper?

*# Efficient Implementations of the Generalized Lasso Dual Path
Algorithm*

In the fused lasso setting the coefficients $\beta \in \mathbb{R}^p$ can be thought of as nodes of a given undirected Graph G , with edge set $E \subset 1, \dots, p^2$. Now lets assume that E has m edges which are enumerated e_1, \dots, e_m . The fused lasso penalty matrix D is then $m \times p$, where each row corresponds to an edge in E . So when $e_l = (i, j)$, we write l th row of D as

$$D_l = (0, \dots, -1, \dots, 1, \dots) \in \mathbb{R}^p,$$

meaning D_l has all zeros except for the the i th and j th location.

Chapter 42

quick summary what I got so far from the algorithm

1.2 case $X=I$. very general dual path algorithm then the use general X , here they compute moore-penrose pseudoinverse and substitute X_{tilde} and y_{tilde} . Also when $p > n$ X does not have full column rank they add a diagonal matrix to the rows, with ϵ time beta magnitude on the diagonals (see github)

implementation: do not solve least square problems all the time, but use QR decomposition in the beginning and then later update the QR decomposition. in some cases more meaningful to take advantage of special structures of D . In our case they use Laplacian linear systems

QR-based general X computing $D_{\text{tilde}} = DX +$ destroy special structures in D , need other methods see Section 5 NOTE: total number of steps are not understood really they use direct solvers because past decisions influence future outcome

2. QR-based general D QR in appendix but role of m and n are changed. two strategies: wide and tall our case tall strategy $DPG=QR$, rotated QR decomposition
3. Special implementation for fused lasso, $X=I$ computing DDT are highly sparse but. are underdetermined because m larger than n , more edges than nodes. We can find arbitrary solution (we will called it basic solution) and from arbitrary solution we can find solution with minimum l_2 norm.
4.1 There is not necessarily a computational improvement but for special cases as for example fused lasso there are improvements because of special structure in D . they compute projection onto the null and basic solutions of linear systems, then from basic solution, optimal solution.

Chapter 43

what to they actually use now

They change algorithm 2 becaus using x_{tild} and stuff because are options are needed when D has structure using x_{tild} will destroy special structure in D . So when X general we use x_{tild} but using x_{tild} will destroy structure in D , therefore when X general AND D has structure use following:

Chapter 44

test lasso new plots

Placeholder

44.1 LASSO model

44.2 Error plots

44.3 Coefficient plots

44.4 Inspect predictions from each fold

44.5 Inspect predictions from best CV-lambda

44.6 Summary