Georgia Weather Analysis

-Gopal Rander A53210491

This is a report on the historical analysis of weather patterns in an area that approximately overlaps the area of the state of Georgia.

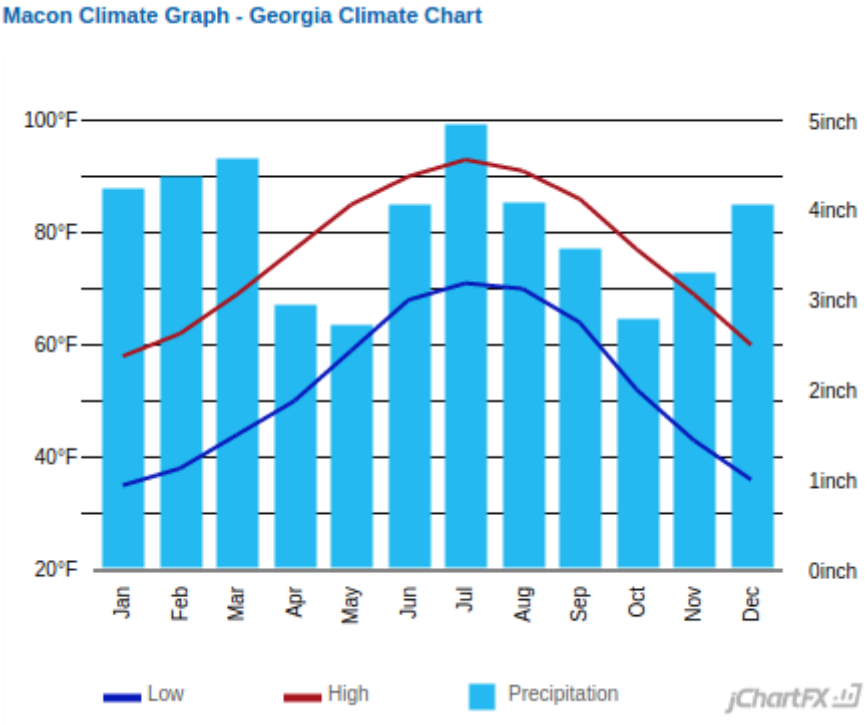
The data we will use here comes from NOAA (https://www.ncdc.noaa.gov/). Specifically, it was downloaded from This FTP site (ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/).

We focused on six measurements:

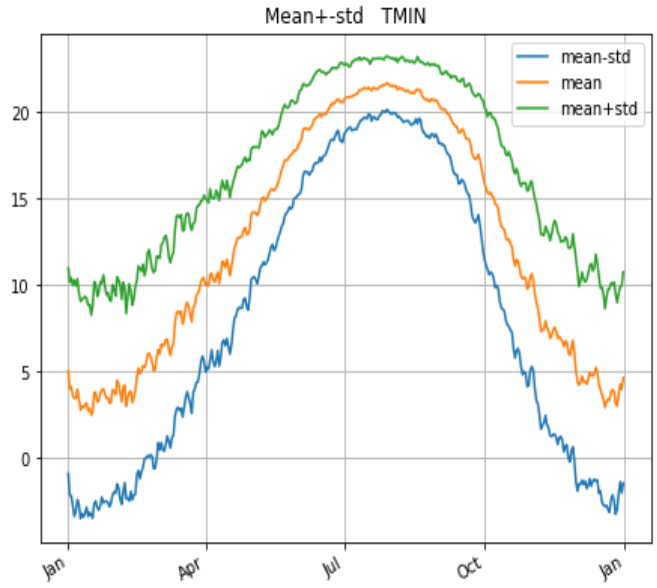
- **TMIN, TMAX:** the daily minimum and maximum temperature.
- **TOBS:** The average temperature for each day.
- **PRCP:** Daily Percipitation (in mm)
- **SNOW:** Daily snowfall (in mm)
- **SNWD:** The depth of accumulated snow.

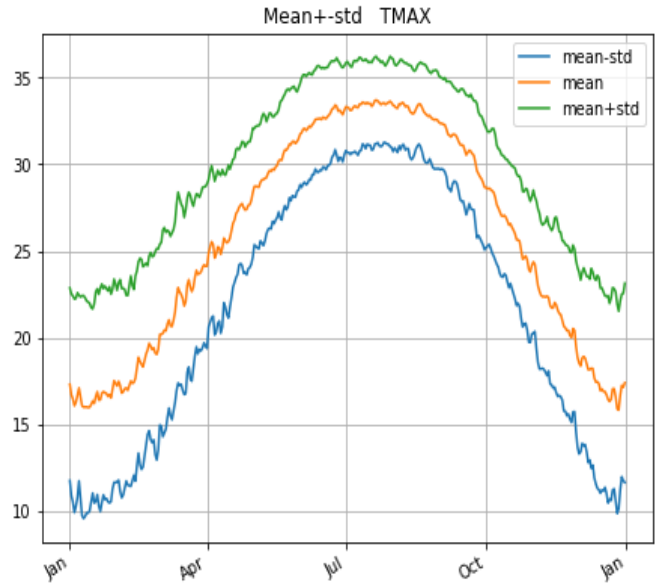
Sanity-check: comparison with outside sources

We start by comparing some of the general statistics with graphs that we obtained from a site called US Climate Data (http://www.usclimatedata.com/climate/macon/georgia/united-states/usga0346) The graph below shows the daily minimum and maximum temperatures for each month, as well as the total precipitation for each month.

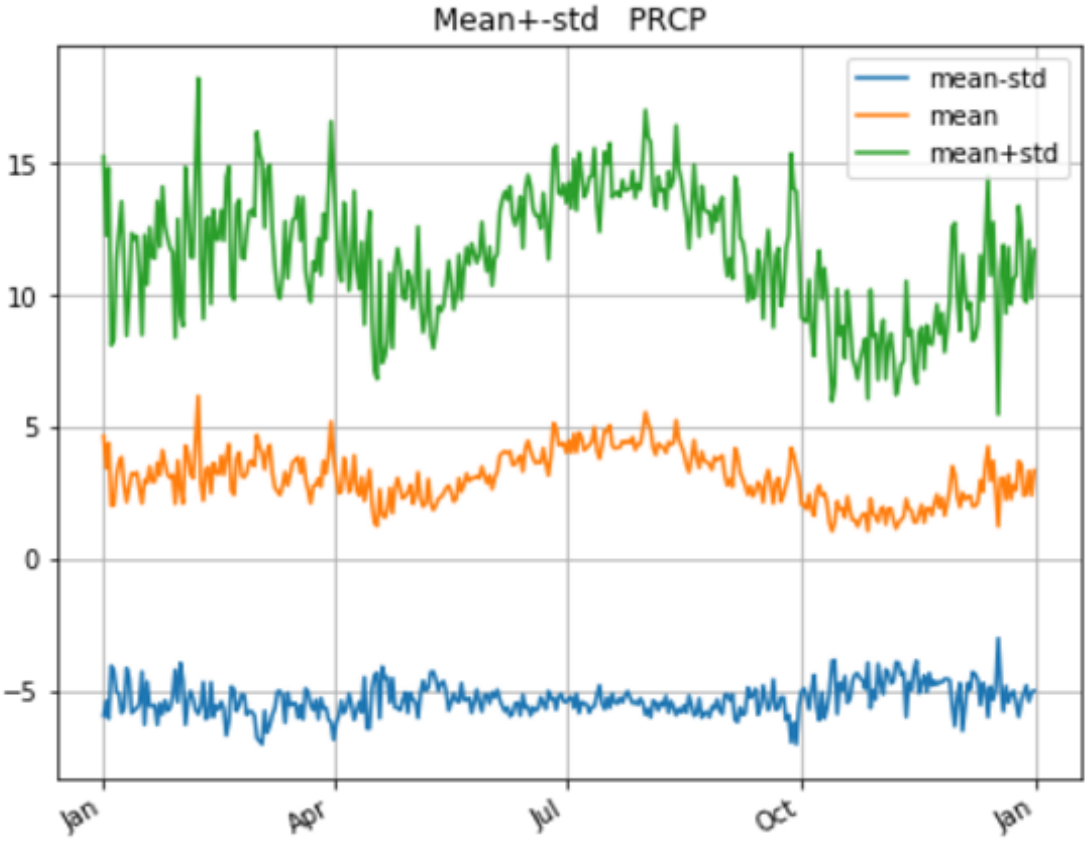


We see that the min and max daily temperature agree with the ones we got from our data, once we translate Fahrenheit to Centigrade.





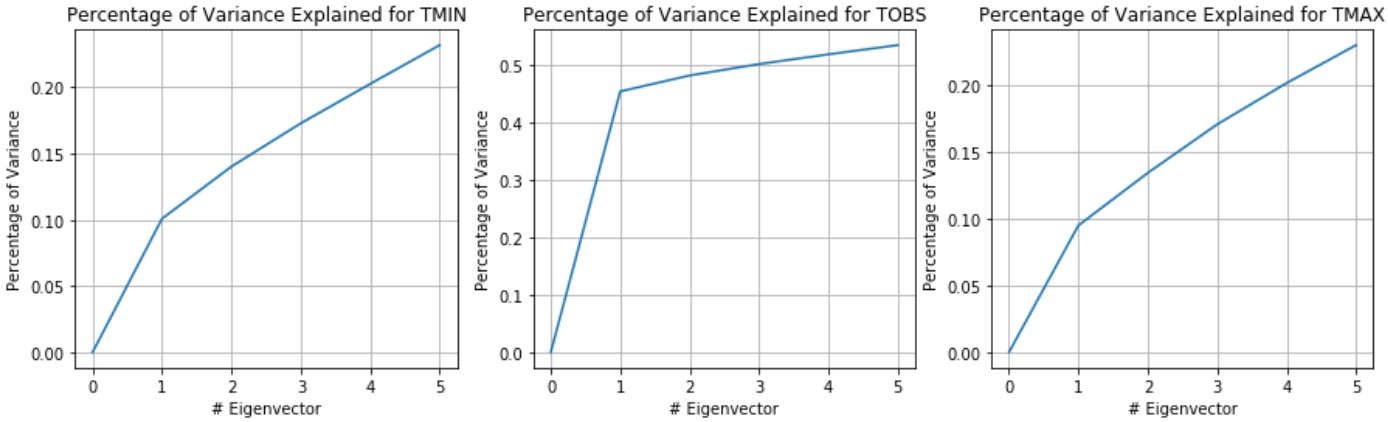
To compare the precipitation we need to translate millimeter/day to inches/month. According to our analysis the average rainfall is 3.00 mm/day which translates to about 3.54 Inches per month. According to US-Climate-Data the average rainfall is closer to 3.8 inch per month. However, there is clear agreement that average precipitation is higher during July, August and dips in October, November which is in-line with the US climate data analysis



PCA analysis

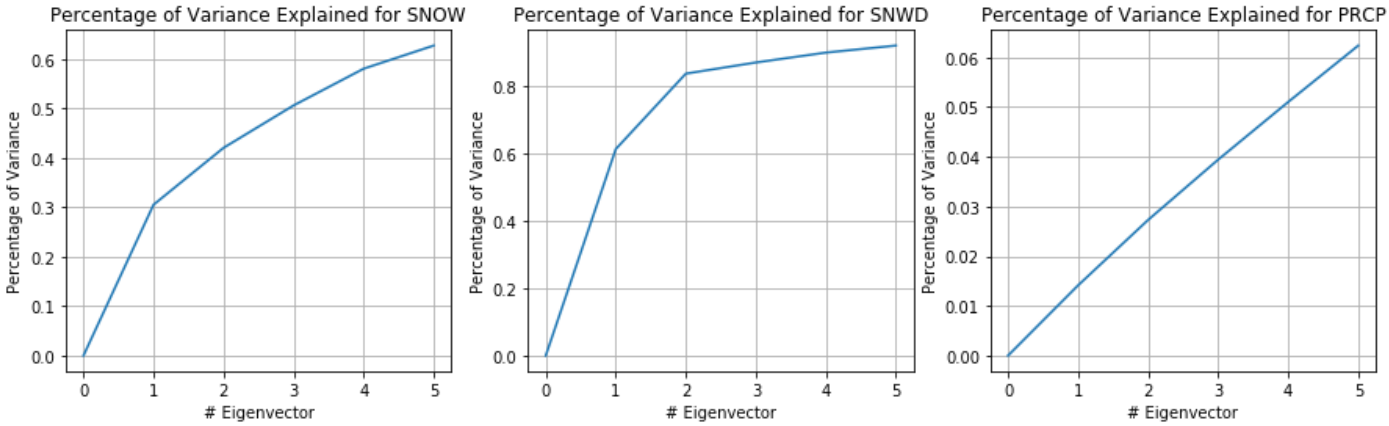
For each of the six measurement, we compute the percentate of the variance explained as a function of the number of eigen-vectors used.

Percentage of variance explained.



We see that the top 5 eigen-vectors explain 23% of variance for TMIN, 55% for TOBS and 23% for TMAX.

We conclude that of the three, TOBS is best explained by the top 5 eigenvectors. This is especially true for the first eigen-vector which, by itself, explains 45% of the variance.



The top 5 eigenvectors explain 6.5% of the variance for PRCP and 62% for SNOW. Both are low values. On the other hand the top 5 eigenvectors explain %90 of the variance for SNWD. This means that these top 5 eigenvectors capture most of the variation in the snow signals. Based on that we will dig deeper into the PCA analysis for snow-depth.

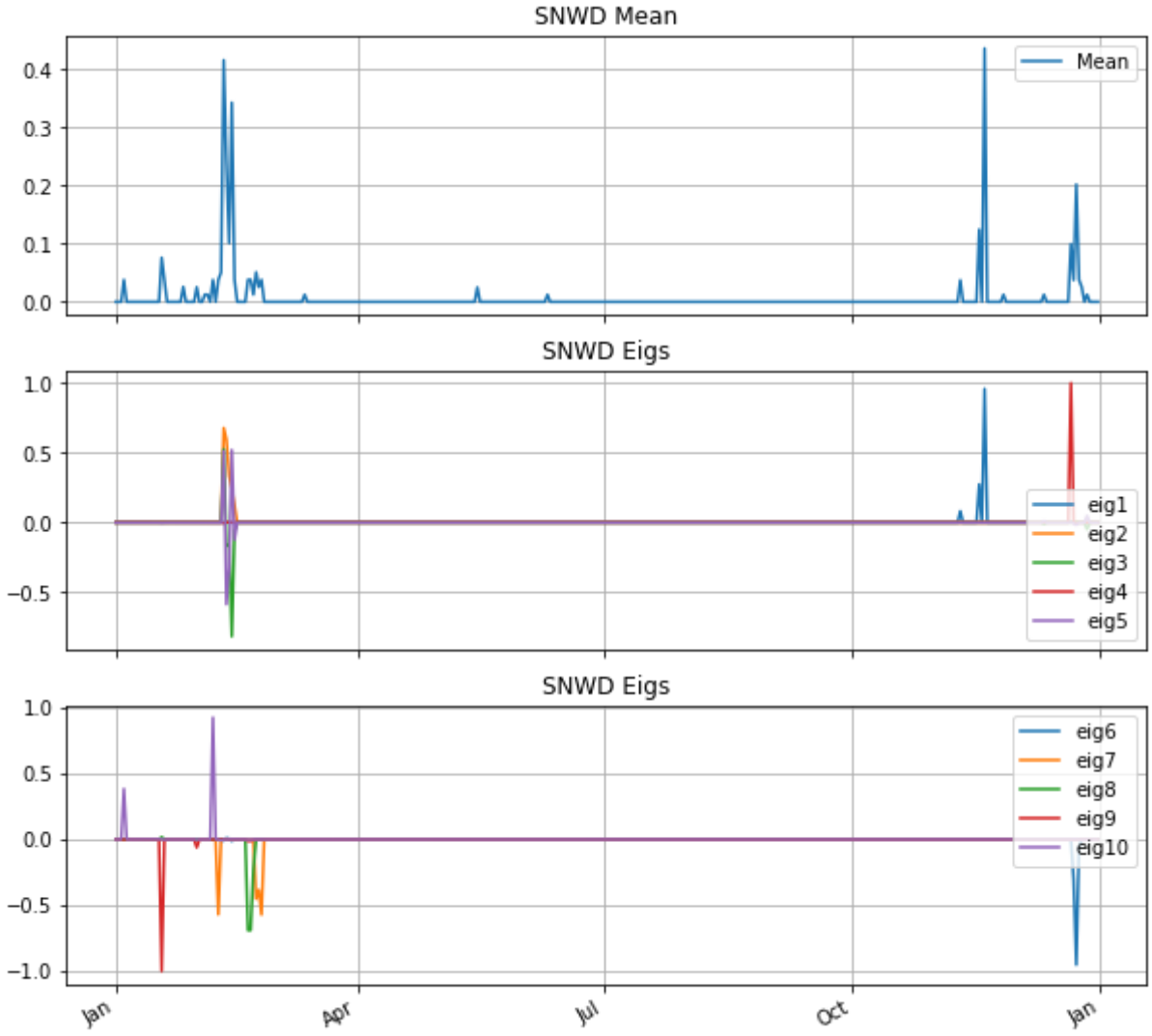
It makes sense that SNWD would be less noisy than SNOW. That is because SNWD is a decaying integral of SNOW and, as such, varies less between days and between the same date on diffferent years. We start by analysis snow depth

Analysis of snow depth

We choose to analyze the eigen-decomposition for snow-depth because the first 10 eigen-vectors explain more than 90% of the variance.

First, we graph the mean and the top 10 eigen-vectors.

We observe that the overall snow depth is very low and sparse. Snow season is during February and March and then during November to December, with very little snow in January. Peak of snow depth is observed in mid-February and mid-November



Next we interpret the eigen-functions. Since the data is very sparse, most of the eigen vectors correspond to the peaks at different time. The first eigen-function (eig1) has peaked at November mid and second eigen function has peaked at February mid. Another observation is that the eigen-function is close to zero during all the times mean is close is zero. The interpretation of this shape is that due to sparse nature of the data, (because there is very less snow in Georgia-Macon area) most of the eigen vectors are similar represenatation of the sparse data. **eig1,eig2 and eig4** are similar in the following way. They all represent the peak at different times. **eig2** captures the dip between peak days of snow.

They can be interpreted as follows:

- **eig1:** snow in nov.
- **eig2 eig 5:** snow in feb, less snow in feb end, slightly more snow in march.
- **eig4:** snow in dec.

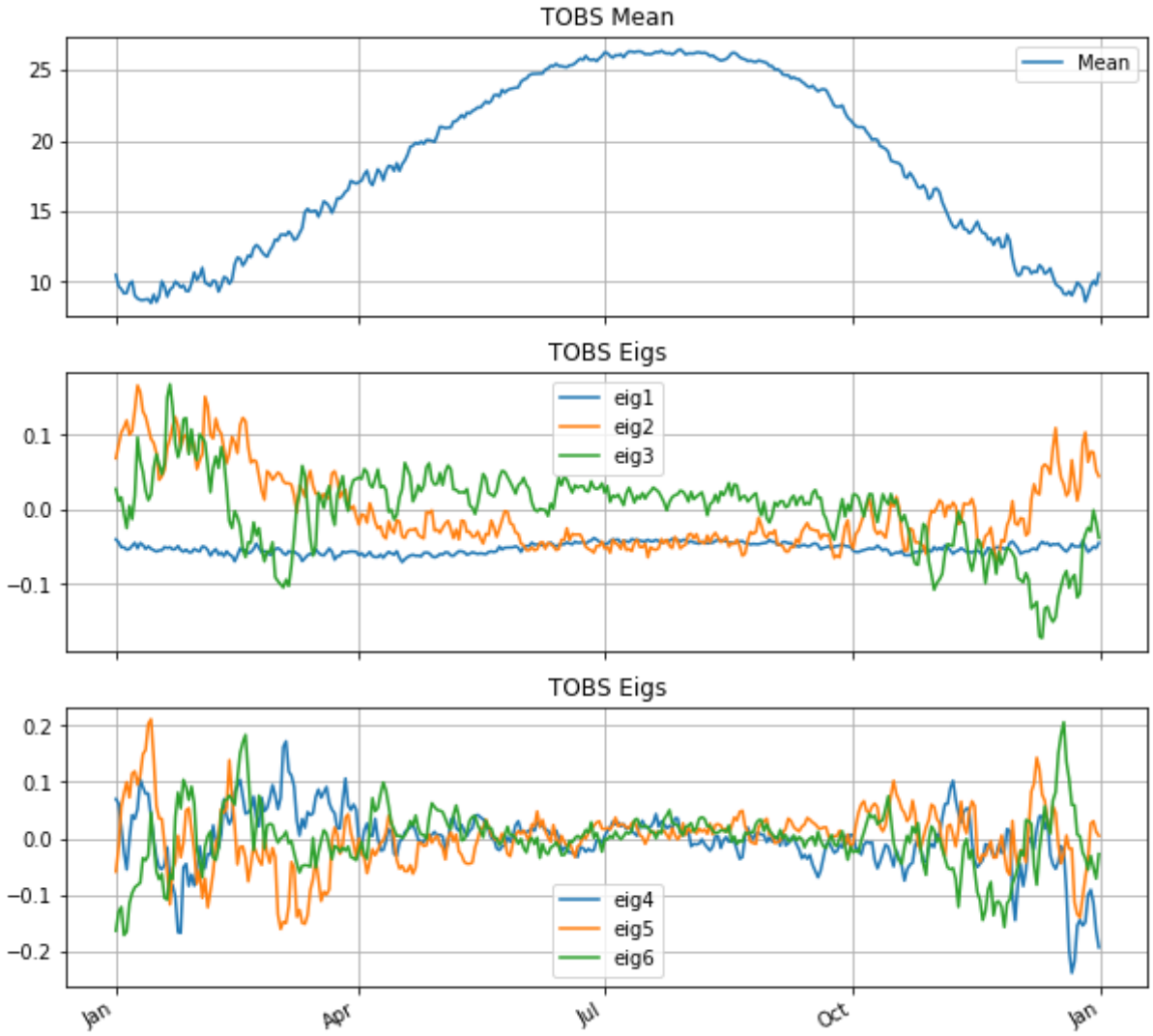
As we observed in the snow depth / snow analysis, the eigen vectors are not rich in information and just a similar represenataion of data as it is. So we decide to analyse TOBS : Average Temp. for a day. Top 5 eigen vectors for TOBS explains close to 55% of variance.

Analysis of Average temperature

We choose to analyze the eigen-decomposition for TOBS (Average temparature each day) because the first 5 eigen-vectors explain 55% variance.

First, we graph the mean and the top 6 eigen-vectors.

We observe that the average temparature follows the seasonal trend i.e. highest in summer during July-August and winter in December and January.

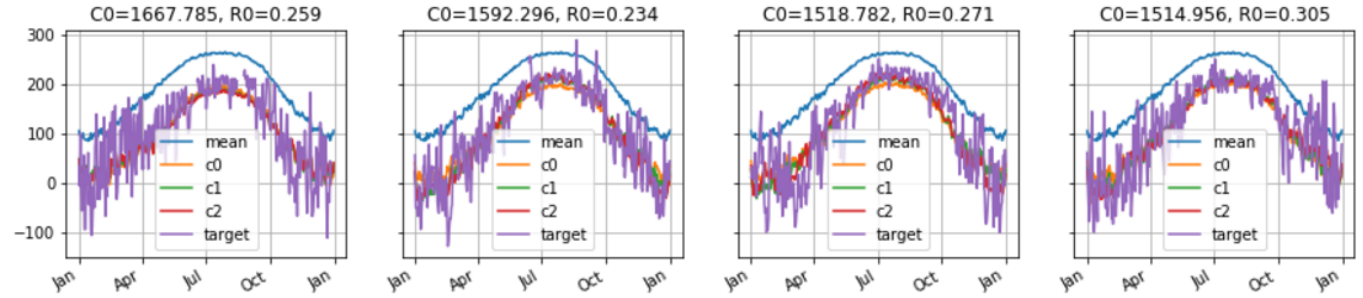


Next we interpret the eigen-functions. Average temparature is less noisy than Max temperature and Minimum temperature. The first eigen-function (eig1) represents a bias value and hence is nearly constant.

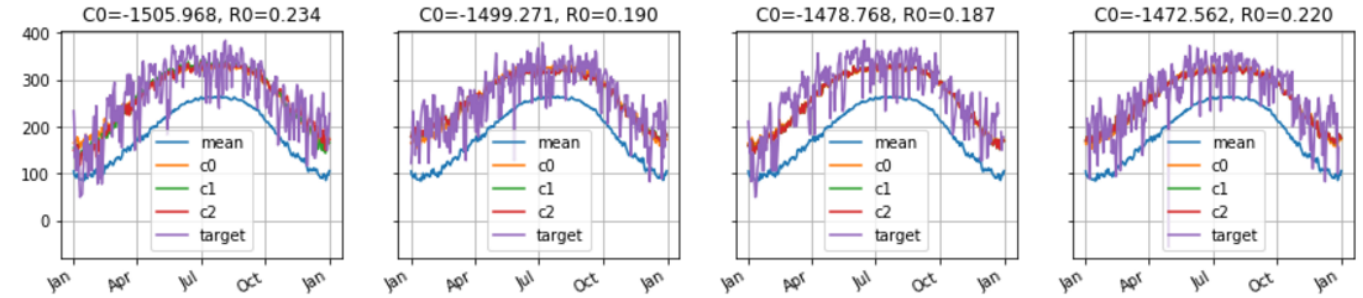
Examples of reconstructions

Coeff0 - corresponding to eig1

Coeff0: most positive



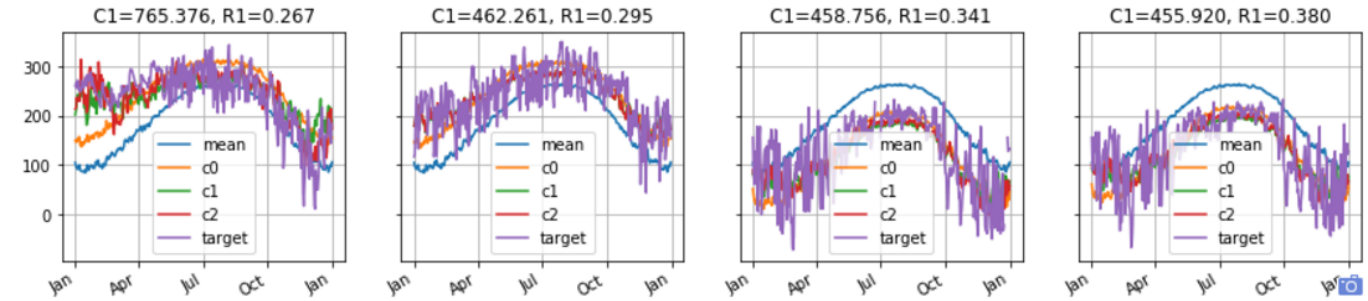
Coeff0: most negative



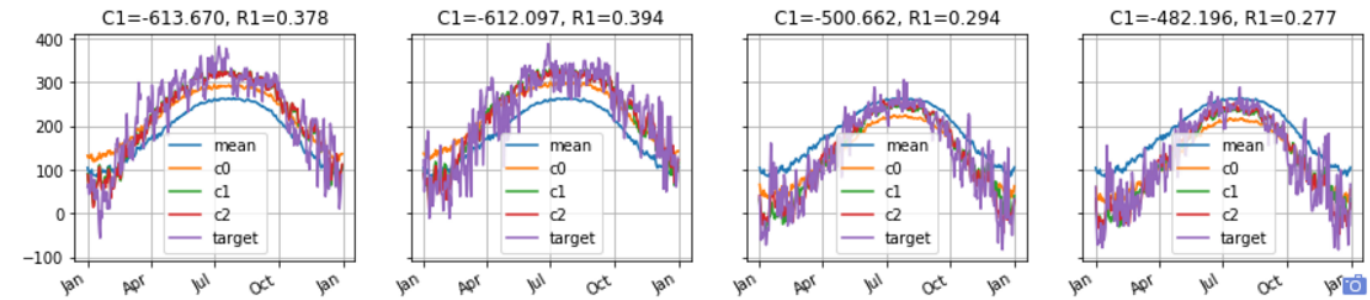
Large positive values of coeff0 correspond to less than average temparature. Low values correspond to more than average temperature.

Coeff1 - corresponding to eig2

Coeff1: most positive



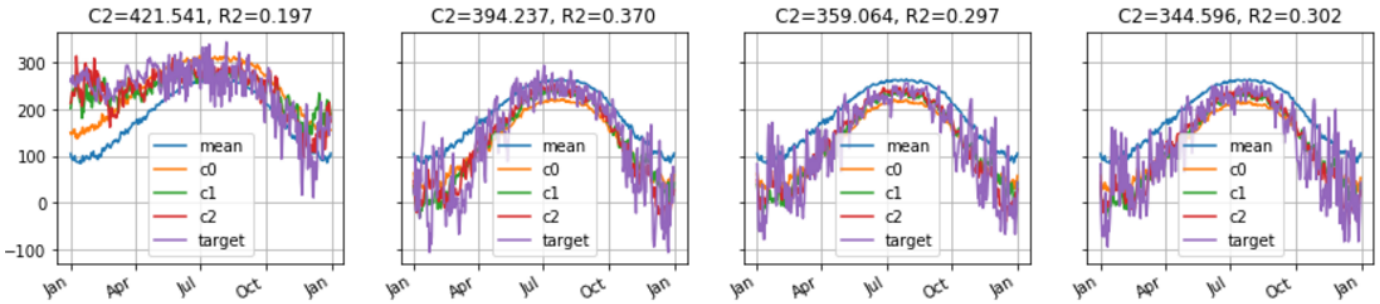
Coeff1: most negative



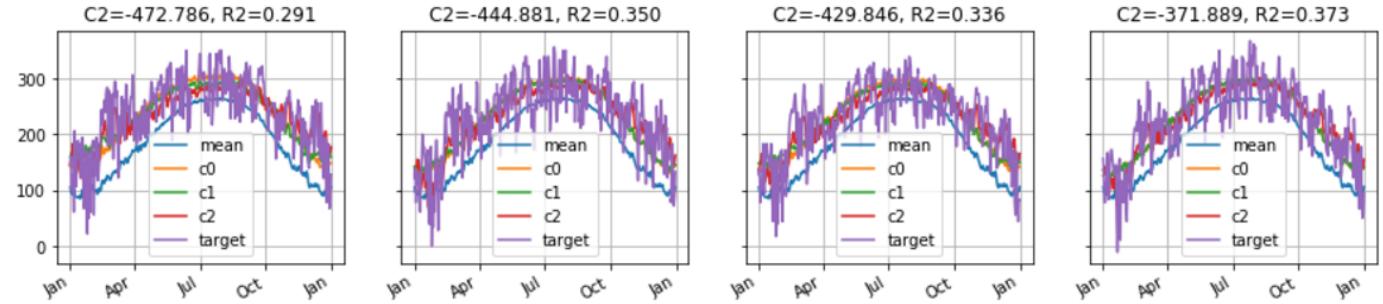
Large positive values of Coeff2 corresponds to more variation in weather during the months of January to April and November to December. Similarly, the most negative values corresponds to cases where the observation is in sync with the mean temperature.

Coeff2 - corresponding to eig3

Coeff2: most positive



Coeff2: most negative



Larger positive values of Coeff_2 corresponds to high temperature during January/February, and then fall in temperature in March/April. Negative values of Coeff_3 corresponds to exactly opposite, i.e. lower temperature in January/February and then rise in temperature during March/April.

The variation in the average temperature of a day is due to both year-to-year variation and station to station variation.

In the previous section we see the variation of Coeff0, which corresponds to the deviation of temperature from mean value, with respect to location. We now estimate the relative importance of location-to-location variation relative to year-by-year variation.

These are measured using the fraction by which the variance is reduced when we subtract from each station/year entry the average-per-year or the average-per-station respectively. Here are the results:

coeff_0
total MS = 612371.43
MS removing mean-by-station= 291151.46, fraction explained=52.45% MS removing mean-by-year = 502487.48, fraction explained=17.94%

coeff_1
total MS = 39001.62
MS removing mean-by-station= 31373.08, fraction explained=19.55%
MS removing mean-by-year = 12996.25, fraction explained=33.32%

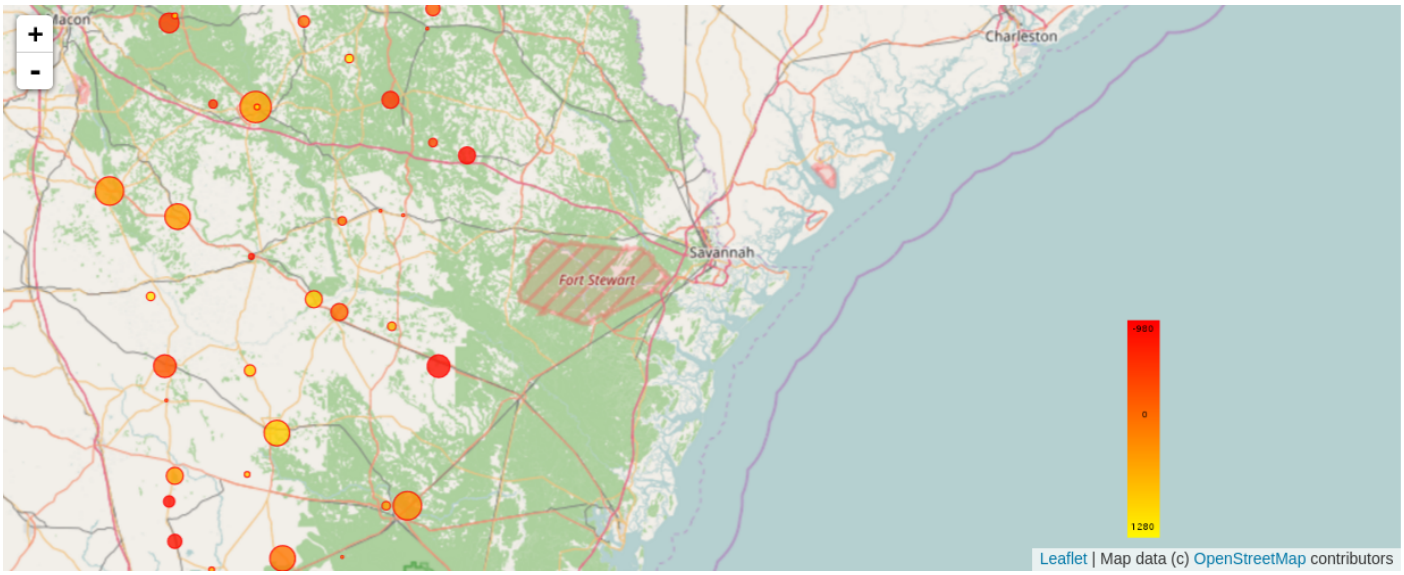
coeff_2
total MS = 26480.89
MS removing mean-by-station= 24856.98, fraction explained= 6.1%
MS removing mean-by-year = 5723.60, fraction explained=78.38%

We see that using coefficent 0, related to eigen vector 1, variation by station explains more than variation by year. Although using coefficient 1 and 2, the fraction explained by variation by year is better than by station. Since coefficient 0 explains 52% of the variation, we conclude that temperature variation is explained by both year-to-year variation and station-to-station variation

Visual representation of TOBS

The map of the staions below shows the number of reading for each station in form of area of circles, and the average value of coefficient 0 (coeff0, corresponding to eigen vector 1) in color coded format.

As stated in report above, red circles show more than mean average temperature, and yellow circles show less than mean average temperature.



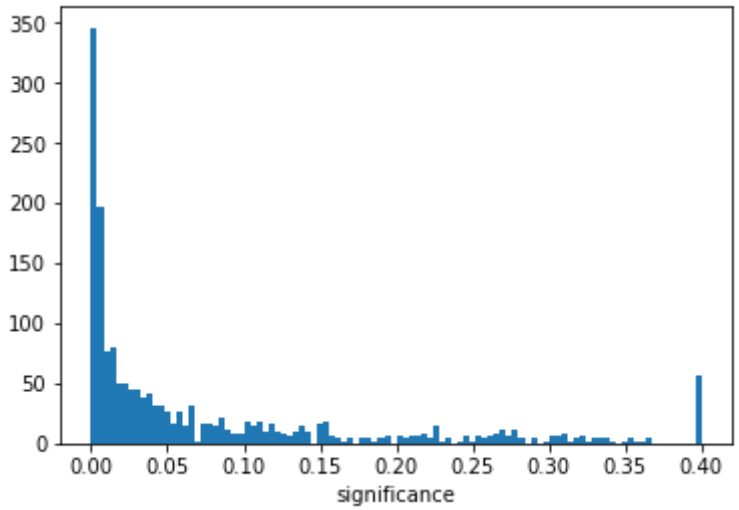
Measuring statistical significance across stations

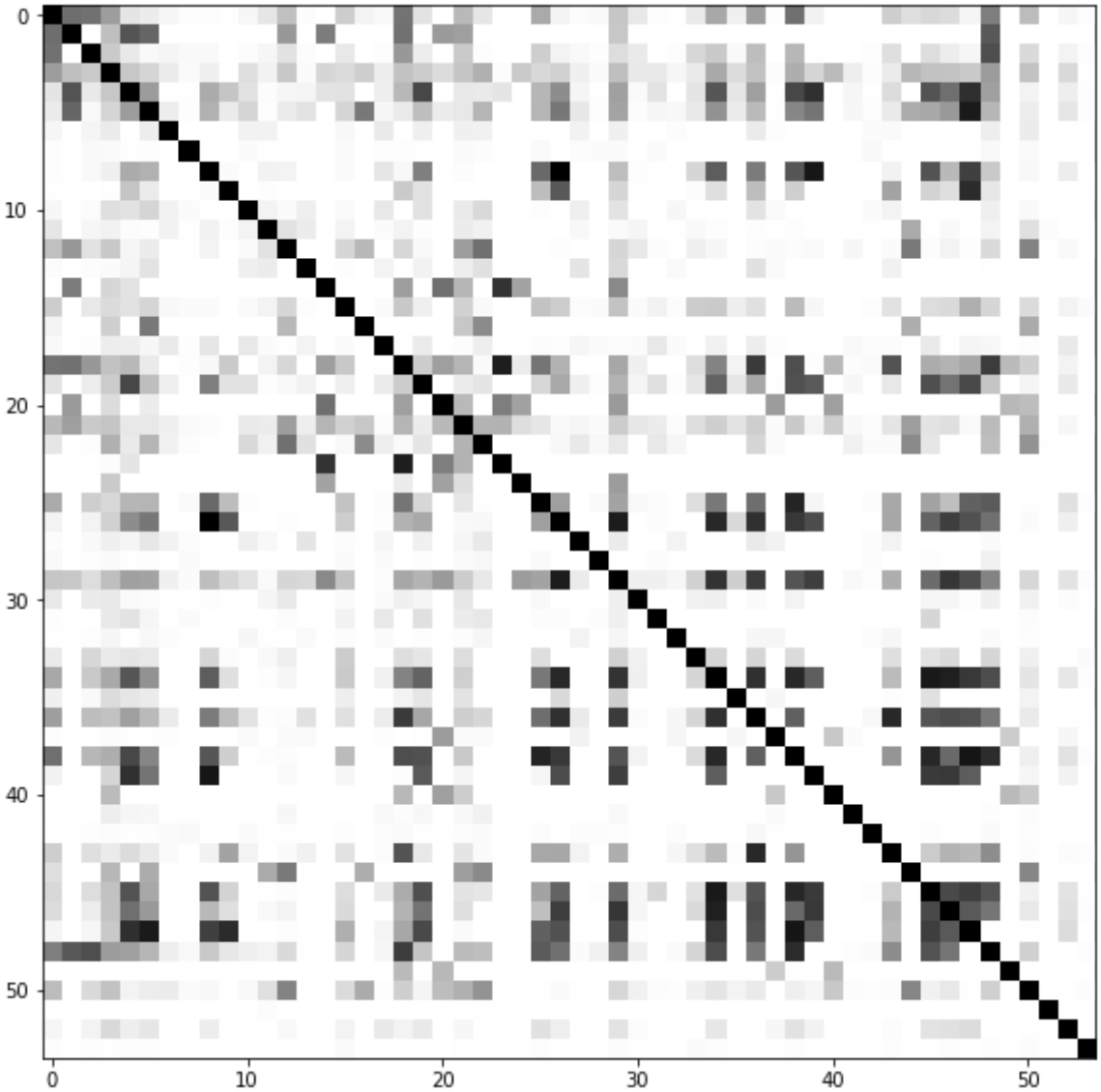
We want to find a statistical test for rejecting the null hypothesis that says that the temperature in the two locations is independent.

Using the inner product is too noisy, because you multiply the temperature on the same day in two locations and that product can be very large - leading to a large variance and poor ability to discriminate.

An alternative is to ignore the absolute temperature, and just ask whether the temperature was greater than a threshold (Eg. 25 degree Celcius). We can then compute the probability associated with the number of overlaps under the null hypothesis.

The histogram below shows the distribution of p-values computed using the above representation. (Ignore the peak at 0.4, it is the same station's p-value with itself). We are using 25 degree celcius as the threshold.


350
300
250
200
150
100
50
0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
significance

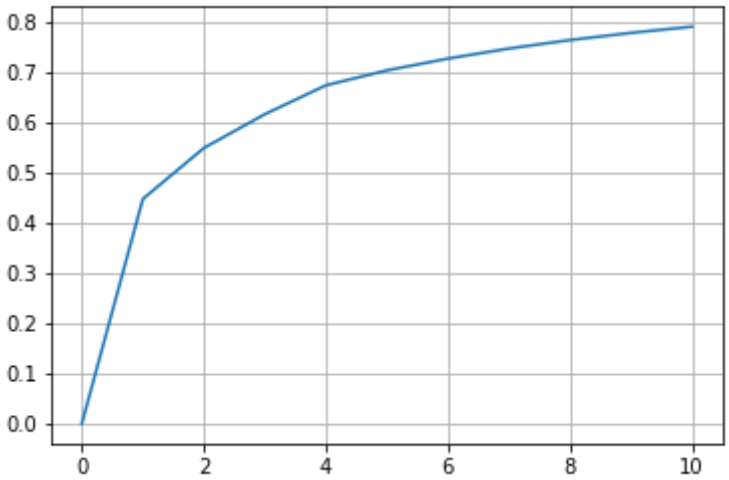

0
10
20
30
40
50
0
10
20
30
40
50

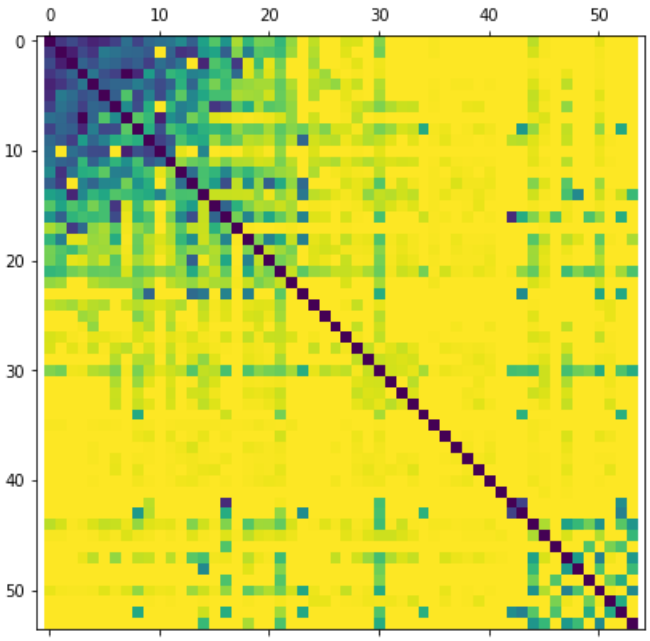
Finding structure in the dependency matrix.

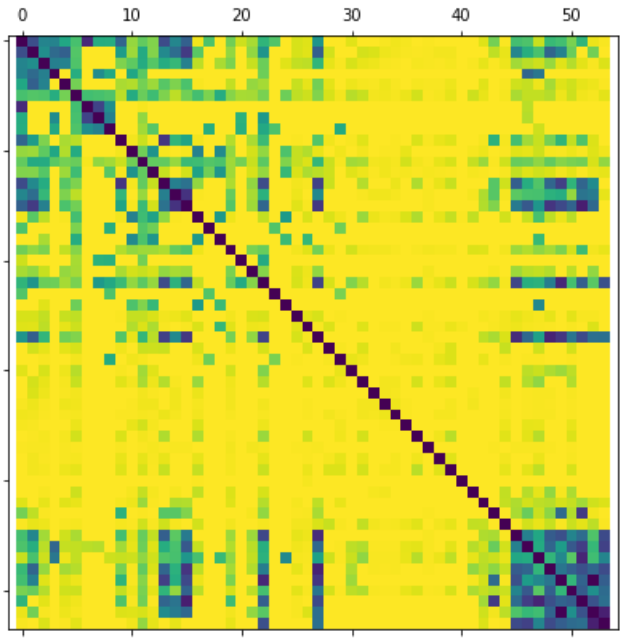
The matrix above shows, for each pair of stations, the normalized log probability that the overlap in high temperature days mostly random. Some station are correlated, which makes sense as temperature on close by stations will not be varying too much. so if a station has > 25 degree temperature, then all close by stations will have a similar reading.

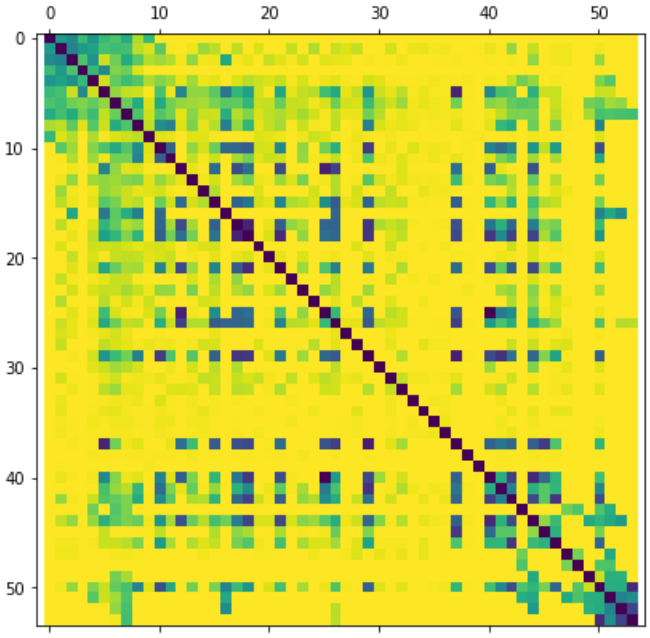
We see immediately the first 3 stations are highly correlatedwith each other. Then a group of 5 stations towards the end are highly correlated.

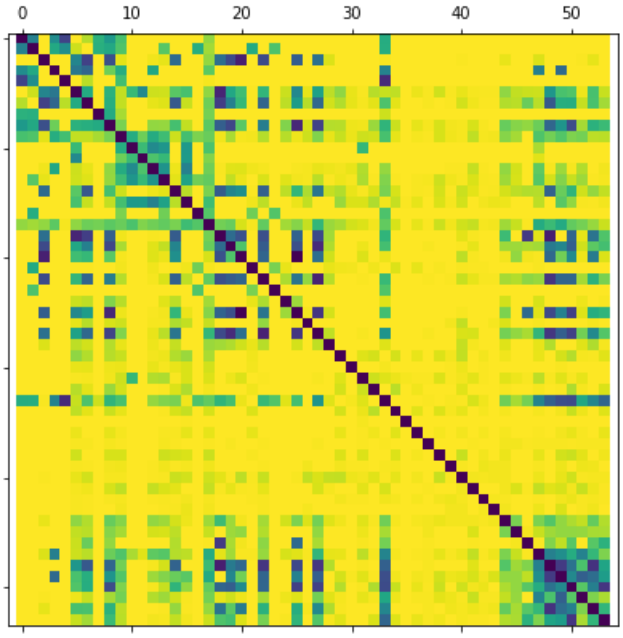
To find more correlations we use SVD (the term PCA is reserved for decomposition of the covariance matrix). As we shall see that the top 10 eigenvectors explain about 80% of the square magnitude of the matrix.


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8
10


0
10
20
30
40
50


0
10
20
30
40
50


0
10
20
30
40
50


0
10
20
30
40
50

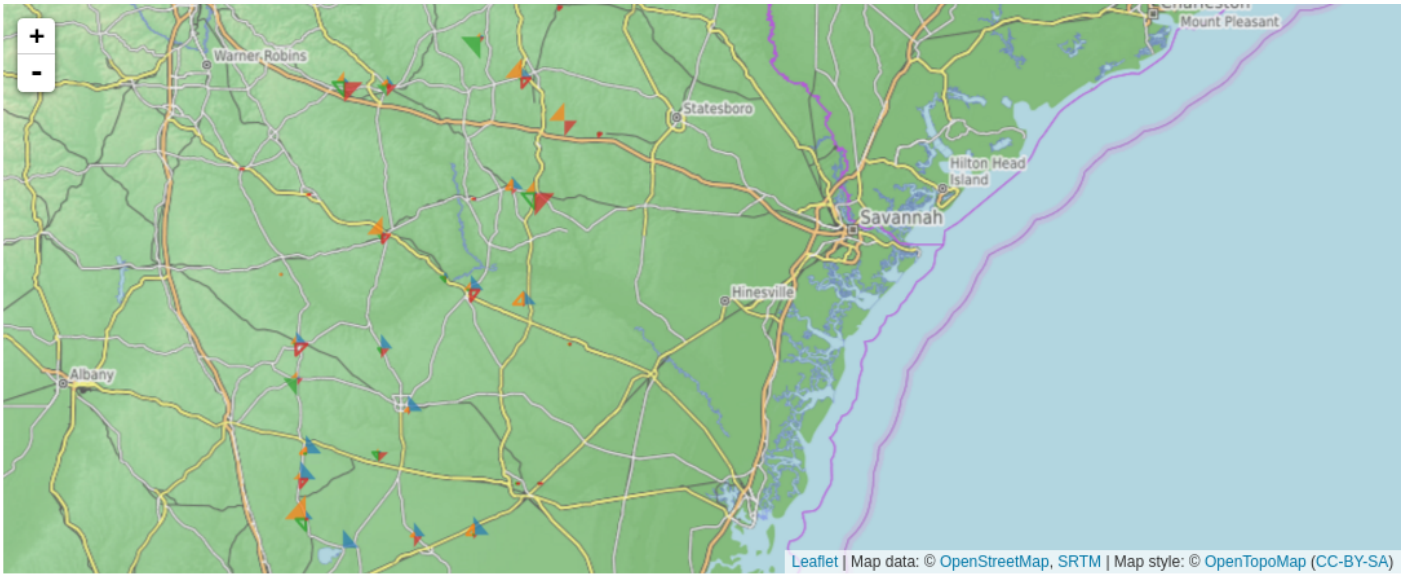
Explanation and possibe extensions

When we reorder the rows and columns of the matrix using one of the eigenvectors, the grouping of the stations becomes more evident. For example, consider the upper left corner of the first matrix (The upper left one). The stations at positions 0-15 are clearly strongly correlated with each other. In second matrix, we see a group of 10 stations co-related.

This type of organization is called **Block Diagonal** and it typically reveals important structure such as grouping or clustering.

Visual Representation of Coefficients for top 4 eigen vectors on above binary representation of data

We represent the coefficient using triangles, with size of triangle as the magnitude of the coefficient. Hollow triangles are negative coefficient and solid are positive. This gives us regional dependency representation of the staions.



We observe that these coefficients are related regionally, which explains the dependency matrix above

In [ ]: