California Weather Analysis

By, Sriram Ravindran - A53208651

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

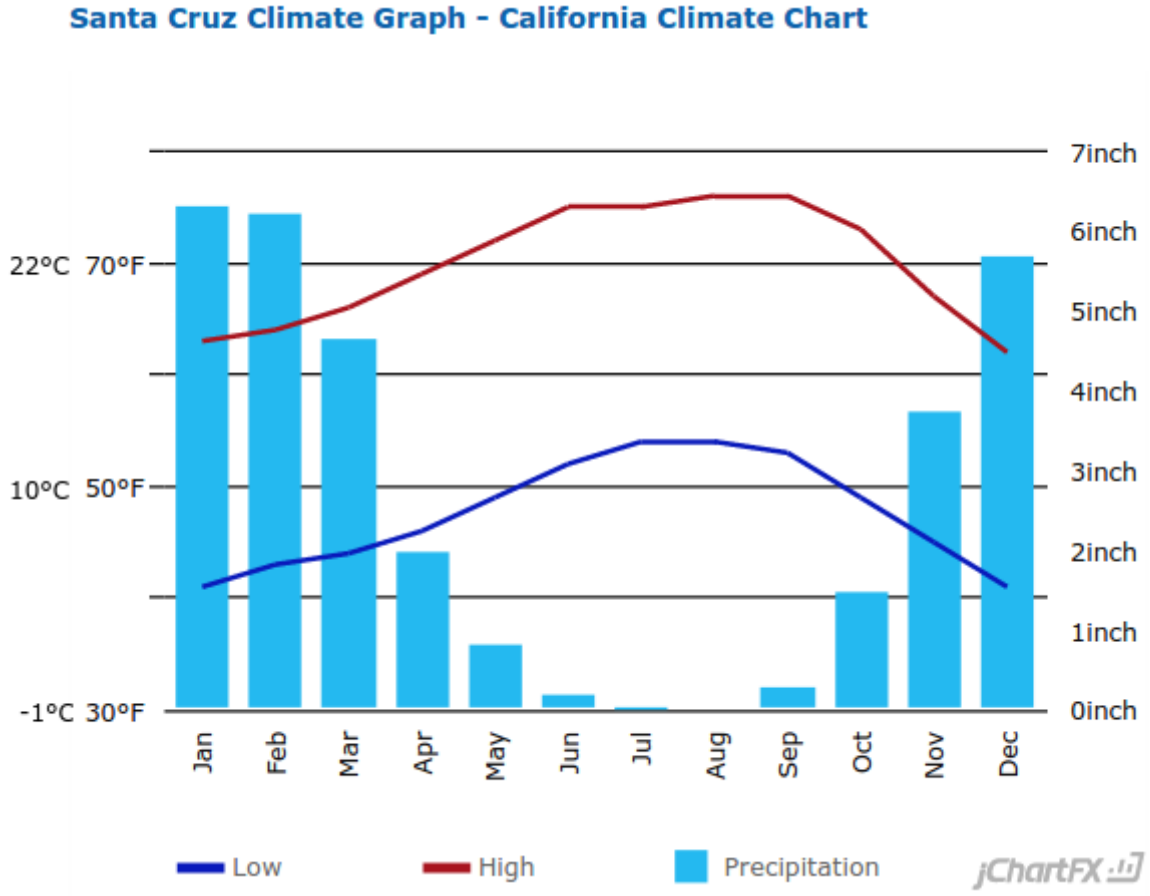
The data we will use here comes from NOAA (https://www.ncdc.noaa.gov/). Specifically, it was downloaded from This FTP site (ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/).

We focused on six measurements:

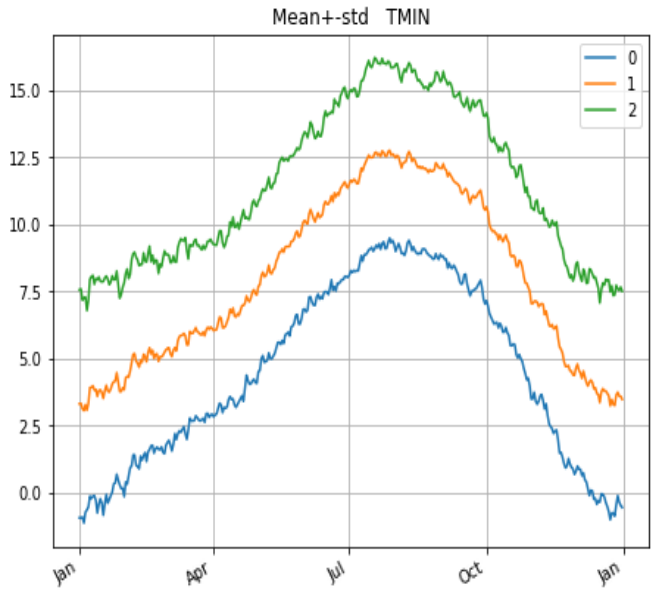
- **TMIN, TMAX:** the daily minimum and maximum temperature.
- **TOBS:** The average temperature for each day.
- **PRCP:** Daily Percipitation (in mm)
- **SNOW:** Daily snowfall (in mm)
- **SNWD:** The depth of accumulated snow.

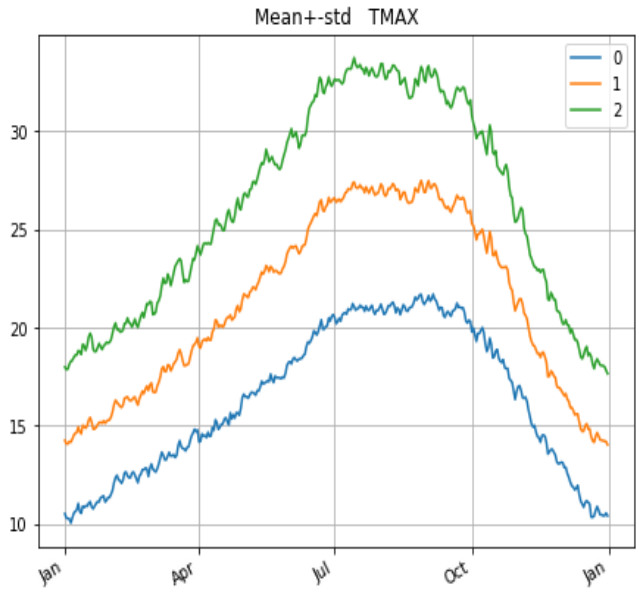
Sanity-check: comparison with outside sources

As done in the sample report, as a sanity check we compare the statistics from a site called US Climate Data (http://www.usclimatedata.com/climate/california/united-states/3174) The graph below shows the daily minimum and maximum temperatures for each month, as well as the total precipitation for each month.

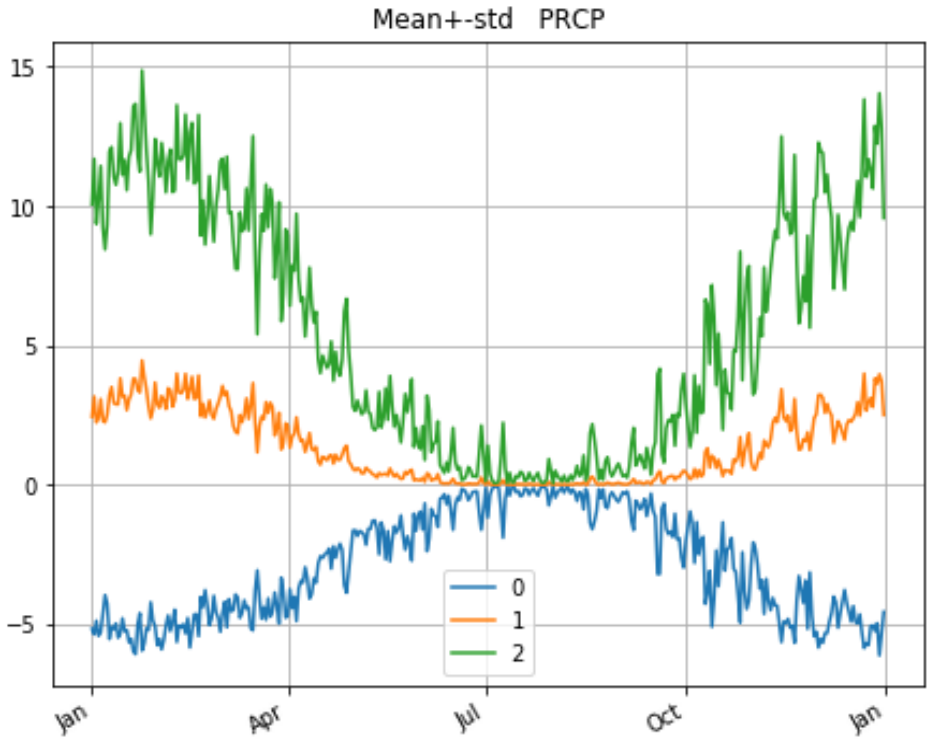


We see that the min and max daily temperature agree within 2-3°C with the ones we got from our data, once we translate Fahrenheit to Centigrade. This is acceptable since our recordings are from multiple regions around San Jose and Santa Cruz (shown).





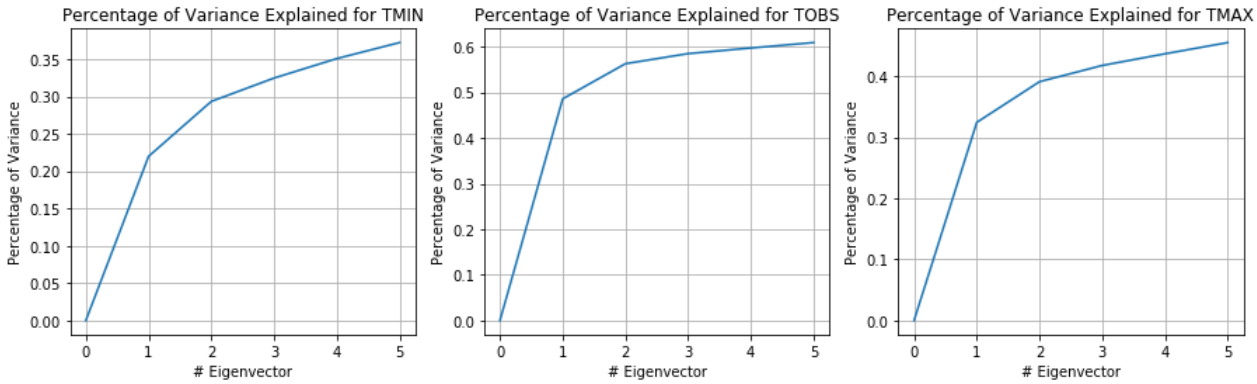
Firstly, we can notice the agreement that it does not rain so much in the summer. Second, we can compare the numbers; according to our data, we have about 485mm of annual rainfall, which is approximately 19inches/year. This is very close to the 18.5in/yr measurement for precipitation for Sacramento according to the US-Climate-Data



PCA analysis

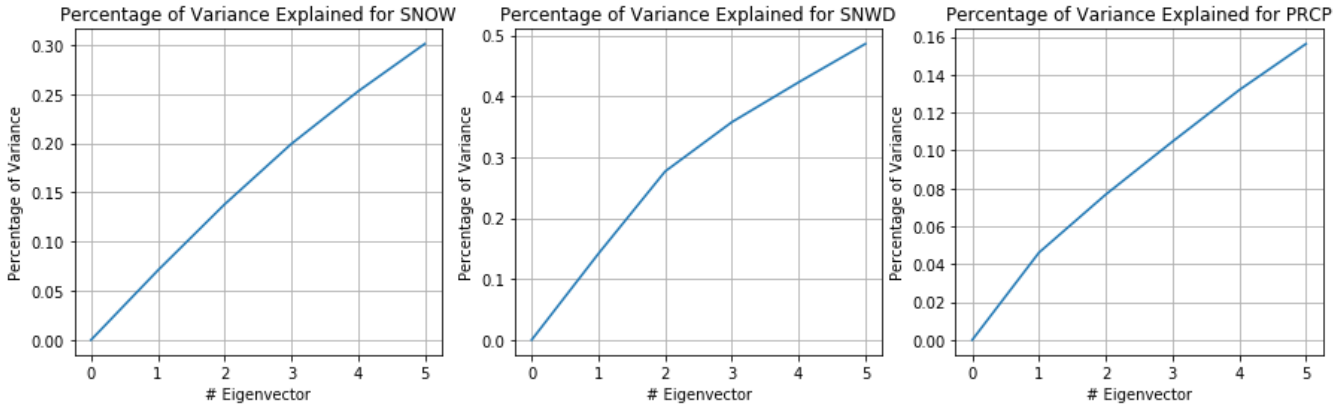
For each of the six measurement, we compute the percentate of the variance explained as a function of the number of eigen-vectors used.

Percentage of variance explained.



We see that the top 5 eigen-vectors explain 37% of variance for TMIN, 61% for TOBS and 45% 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 50% of the variance.



The top 5 eigenvectors explain 16% of the variance for PRCP and 30% for SNOW. On the other hand the top 5 eigenvectors explain 50% of the variance for SNWD. This means that these top 5 eigenvectors capture half of the variation in the snow signals. Based on that we will dig deeper into the PCA analysis for snow-depth.

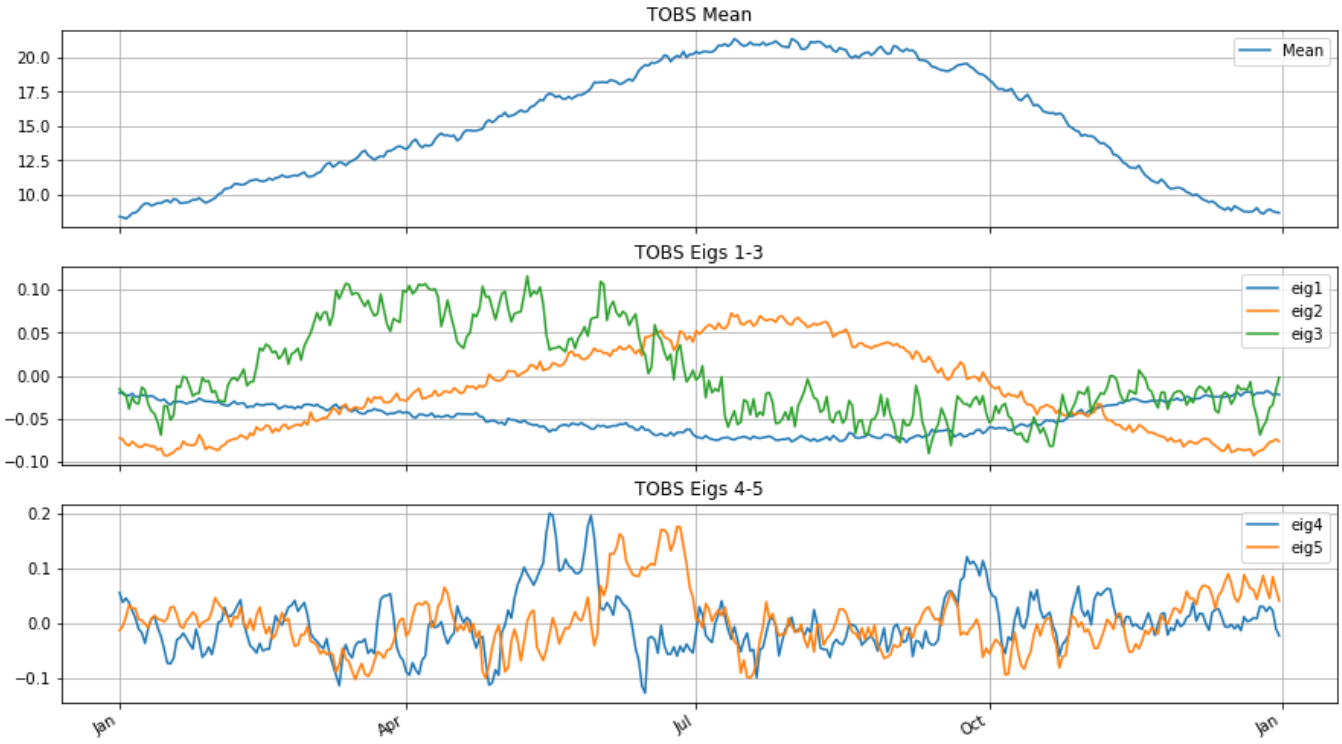
It makes sense that SNWD would be less noisy than SNOW. That is because SNWD is a decaying integral of SNOW and, as such, varies less between days and between the same date on different years.

Analysis of TOBS

We choose to analyze the eigen-decomposition for TOBS because the first 5 eigen-vectors explain 61% of the variance.

First, we graph the mean and the top 5 eigen-vectors.

From the mean curve, we can see that the temperature in the Mat - Oct period which can be considered as summer. T



Next we interpret the eigen-functions. The first eigen-function (eig1) has a shape very similar to the mean function. The main difference is that the eigen-function is close to zero during october-december while the mean is not. The interpretation of this shape is that eig1 represents the overall amount of snow above/below the mean, but without changing the distribution over time.

Next, we move onto the interpretation for the eigen-functions.

eig1 seems to be opposite inverted as the mean curve, it is wholly negative. Depending on the sign and magnitude of the coefficient, it will act as a positive or negative offset as it seems to be mostly flat. Since this is the **first** eigenvector, we might say that TOBS across areas of measurement do not follow the same temperature patterns.

eig2 is very similar in shape to the mean function. It is high in the summer and low in the winters.

eig3 says the average temperature shows an increasing trend till summer then a decreasing trend until winter.

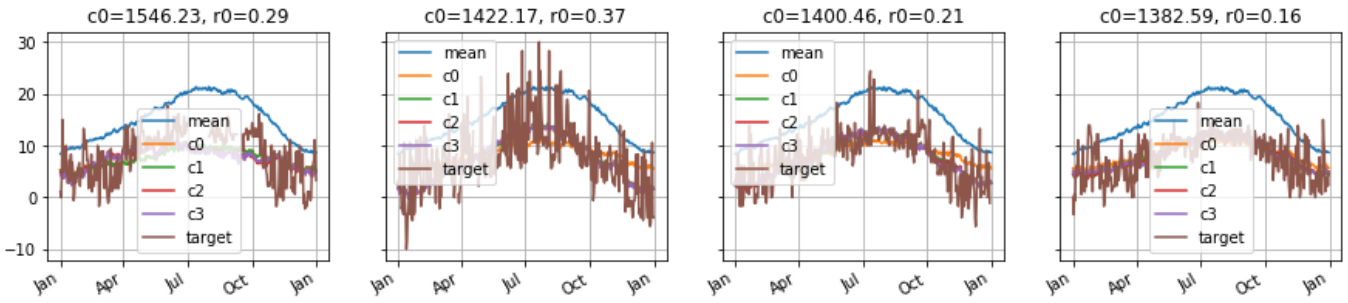
eig4 and eig5 keep oscillating throughout the year. They are capturing the finer details of the distribution, however they have low coefficients and do not affect values that much. They are not very easily interpretable but both seem to peak between April-Jul.

Examples of reconstructions

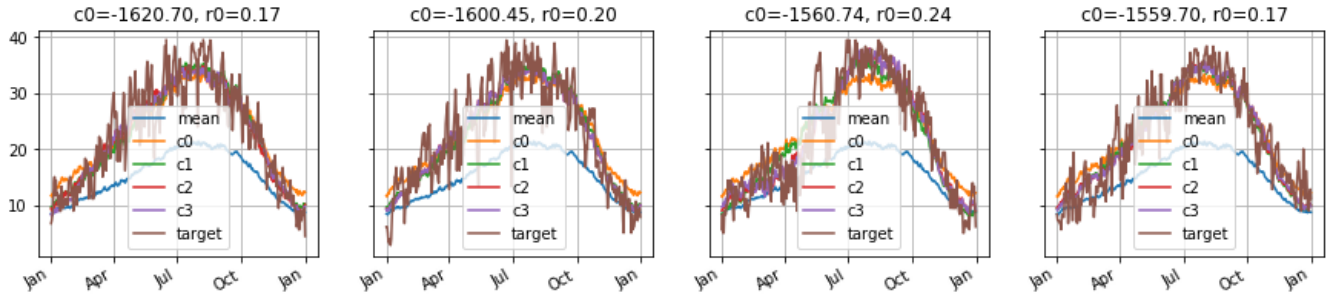
Eig4 and Eig5 capture finer details and are not easily interpretable, therefore they are not shown here.

Coeff0 [df3.res_1 < 0.5]

Coeff0: most positive



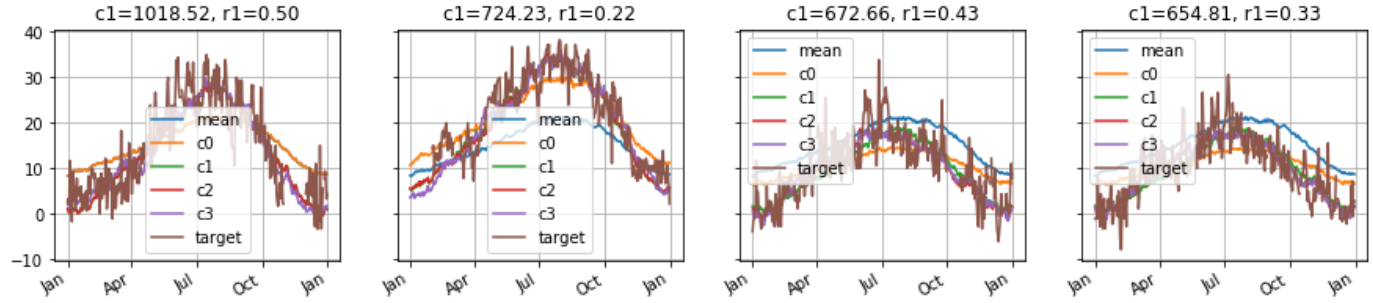
Coeff0: most negative



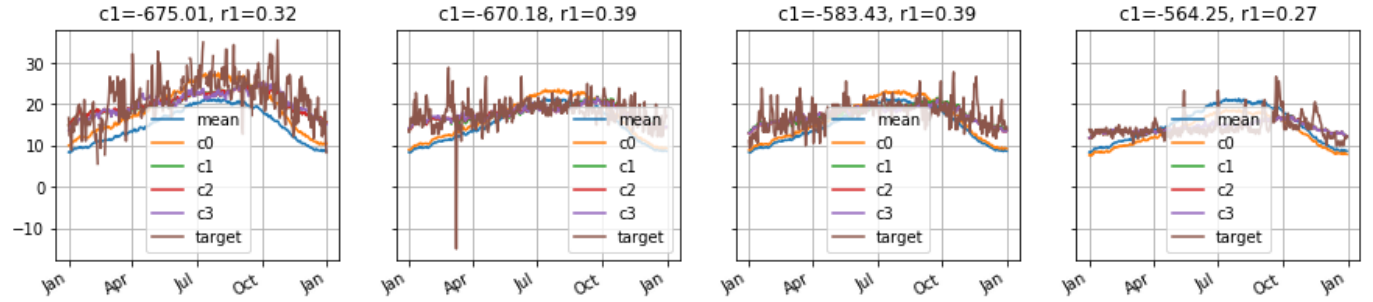
Large negative values of coeff0 correspond to more than average temperature. Large positive values correspond to less than average temperature.

Coeff1 [df3.res_2 < 0.5]

Coeff1: most positive



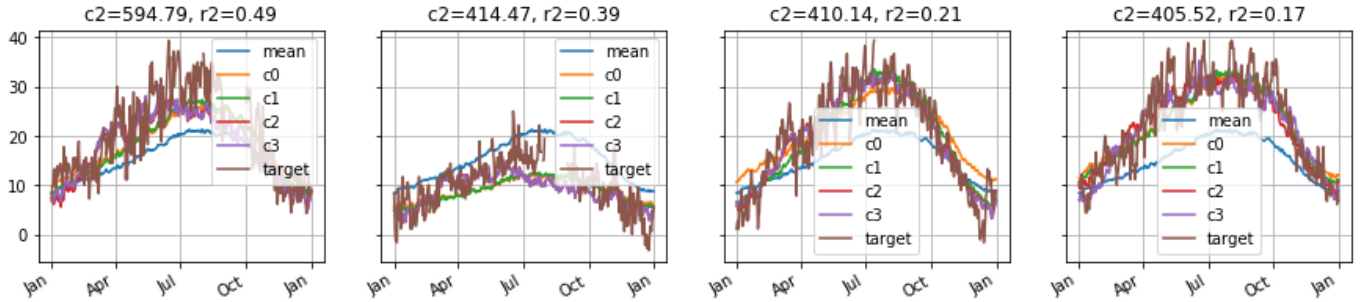
Coeff1: most negative



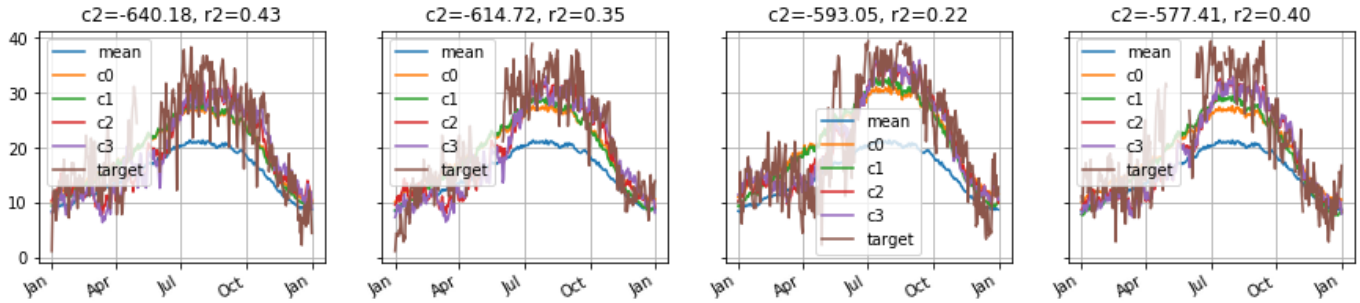
Large positive values of coeff1 correspond to a high temperature in the April-October months and lower temperatures in the rest of the year. With lower coeff0, this effect is less pronounced.

Coeff2 [df3.res_3 < 0.5]

Coeff2: most positive



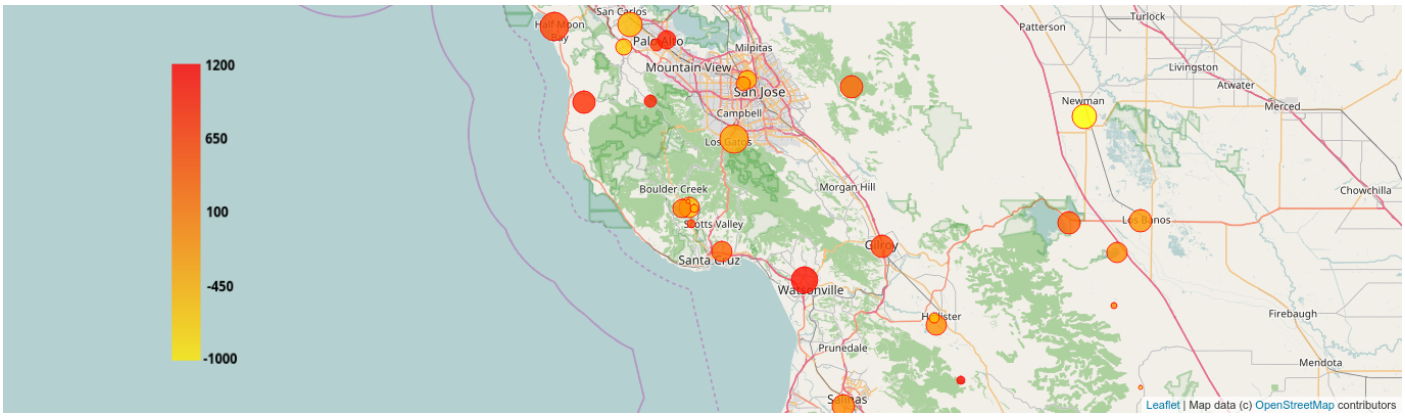
Coeff2: most negative



Large positive values of coeff2 correspond to a temperature trend that increases towards April-October and decreases then onwards. This is consistent with the fact that California is in the northern hemisphere.

Visual representation of spatial variation of TOBS

I have plotted the average temperature on a map using ipyleaflet. The legend shows values of the coef_0. As coef_0 increases, temperature drops. **The more yellow the circles, the higher the temperatures. The red circles are lower temperatures.**



The variation in the timing of TOBS is mostly due to station-to-station variation

In the previous section we see the variation of Coeff_0, which corresponds to the total amount of TOBS. 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 = 499990.58
MS removing mean-by-station= 144991.43, fraction explained=71.1
MS removing mean-by-year = 446552.93, fraction explained=10.68

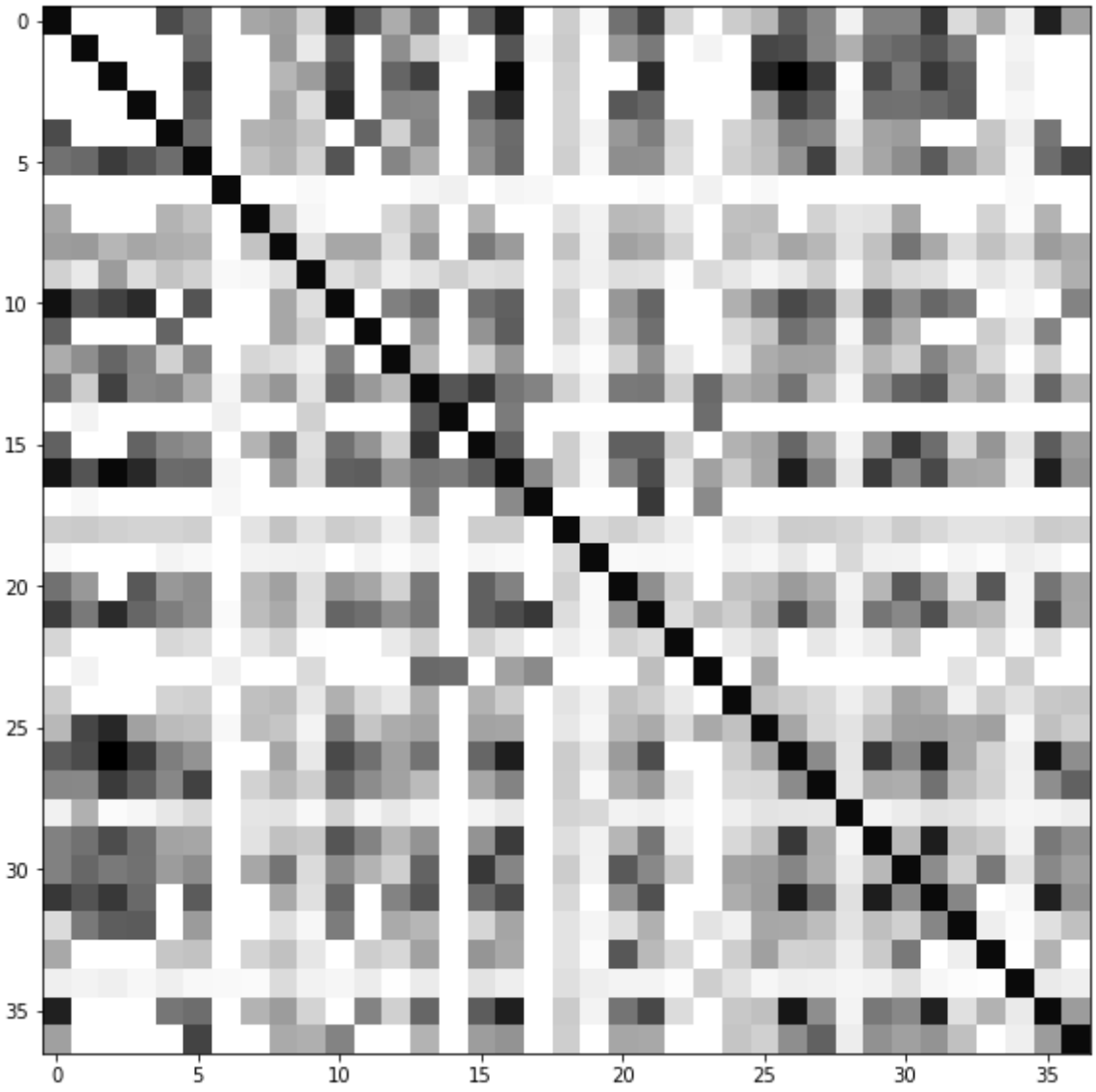
coeff_1
total MS = 81843.78
MS removing mean-by-station= 16837.45, fraction explained= 79.4
MS removing mean-by-year = 71296.66, fraction explained=81.4

coeff_2
total MS = 27995.66
MS removing mean-by-station= 21552.13, fraction explained= 23.01 MS removing mean-by-year = 17355.48, fraction explained=38.0

We see that variation of coef_0 and coef_1 is explained more by variation by station whereas for coef_2 it is explained more by variation by year.

Dependency matrix for TOBS measure between different stations

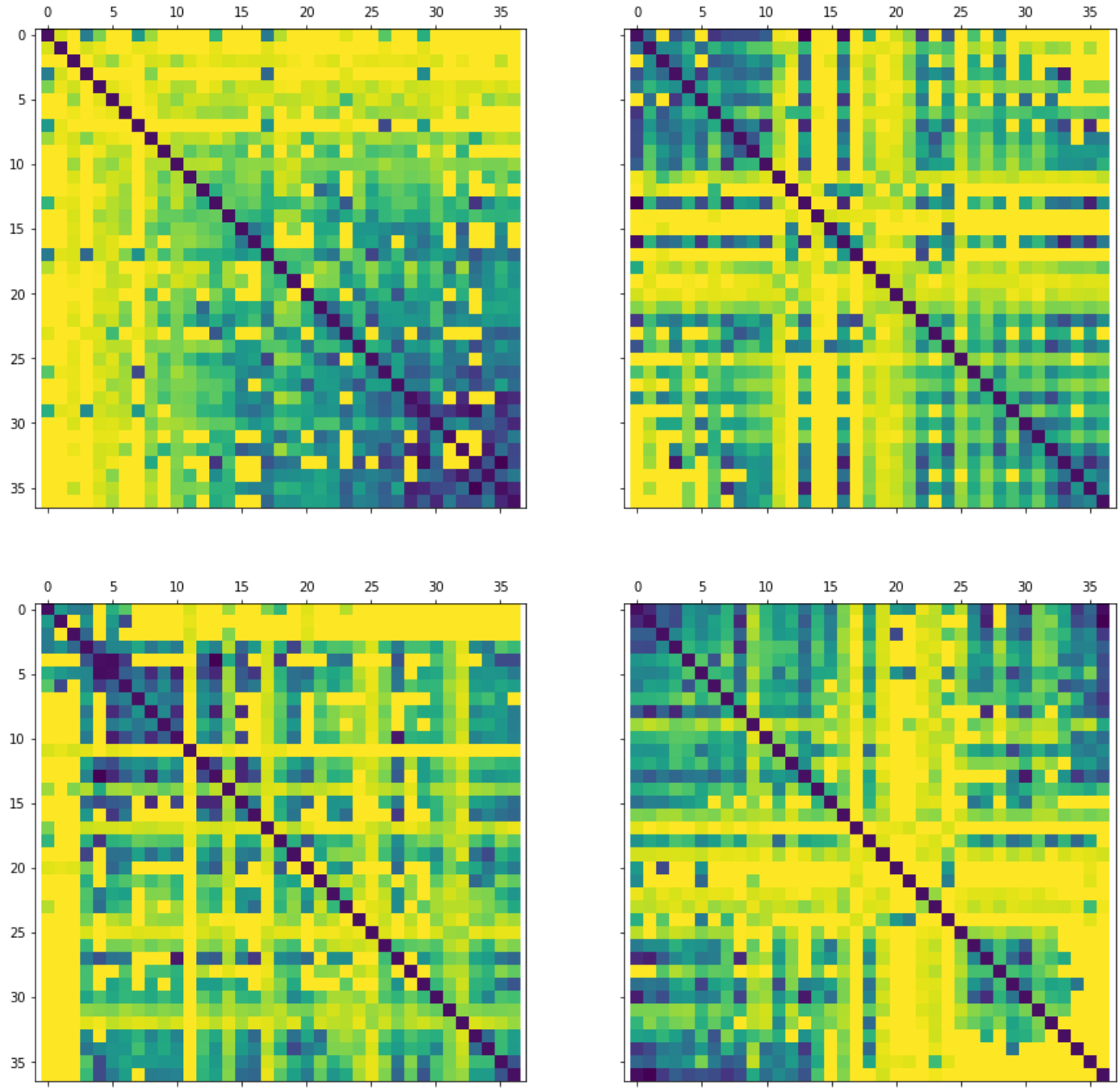
We plotted the dependency matrix for the TOBS measure between different stations. To obtain the graph, we thresholded the temperature at 15°C. This means, any temperature less that that was assumed as 0, any more as 1 (relatively warmer). We obtained the following image.



We can see that a lot of the stations are highly correlated with one another, but a pattern in the image is not clearly discernible. The darker the cell, the higher the correlation. A group of very correlated stations is: USC00043714, USC00040673, USC00040677, USC00049792, USC00047807, USC00040674, USC00047821, USC00044555.

Block chain diagram

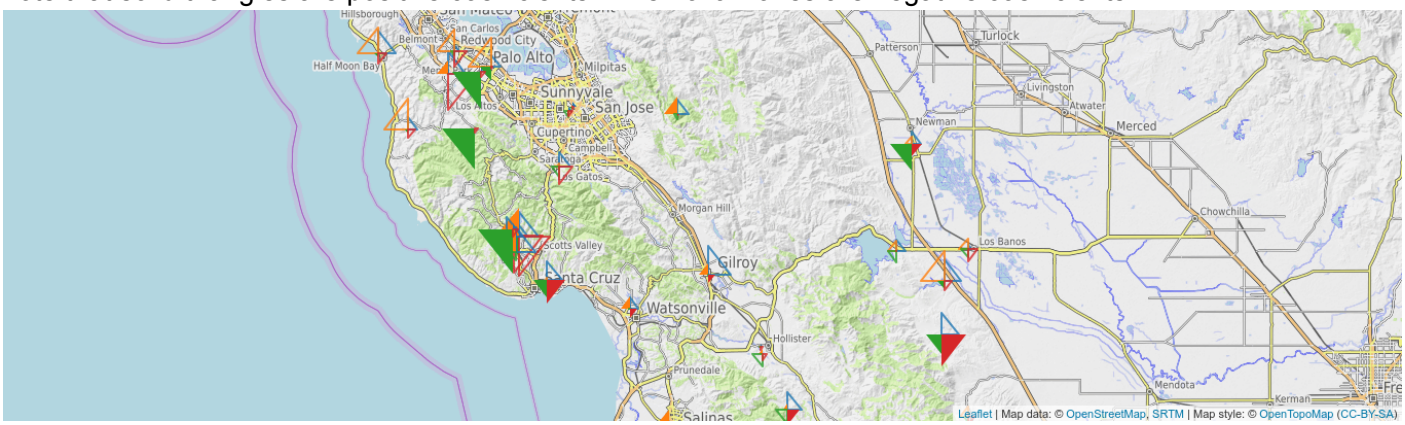
When we reorder the rows and columns of the matrix using one of the eigenvectors, the grouping of the stations becomes more evident. For example, consider the upper left corner of the scond matrix (The upper left one). The stations at positions 0-22 are clearly strongly correlated with each other. Even though there are some stations, in positions 15-18 or so, which are more related to each other than to the rest of this block. This type of organization is called Block Diagonal and it typically reveals important structure such as grouping or clustering.



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 matrix. The stations at positions 25-35 are clearly strongly correlated with each other. Similarly, in upper right diagram, the top left corner reveals that 0-10 stations are highly correlated.

Geographical representation of correlation between stations on TOBS (0 if < 15°C else 1)

This was generated using the first 4 eigenvectors. Each eigenvector has an associated color between green, blue, orange and red. The size of the triangle at each station indicates the strength of the coefficent for that eigenvector. Thus, if two stations have similar prominent colored triangles, they are similar to one another. Do note that solid triangles are positive coefficients while hollow ones are negative coefficients.



In [ ]: