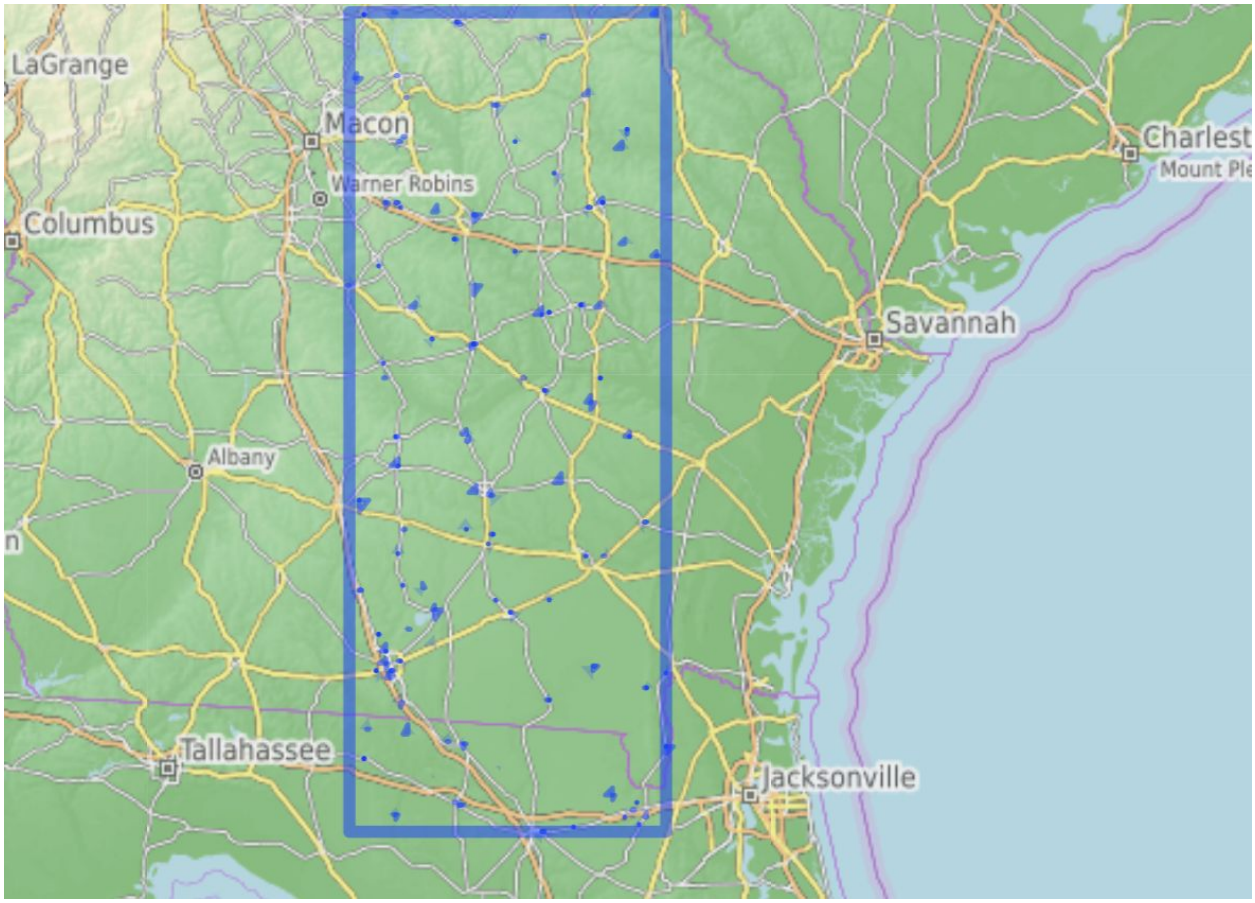
CSE 255: Homework 5

Analysis of Weather Data: Georgia Region (BSBSSSBB)

Shriram Kumar (PID : A53221613)

Introduction

The area code BSBSSSBB corresponds to a region in the state of Georgia. The map below shows the region with the stations where weather was recorded marked.

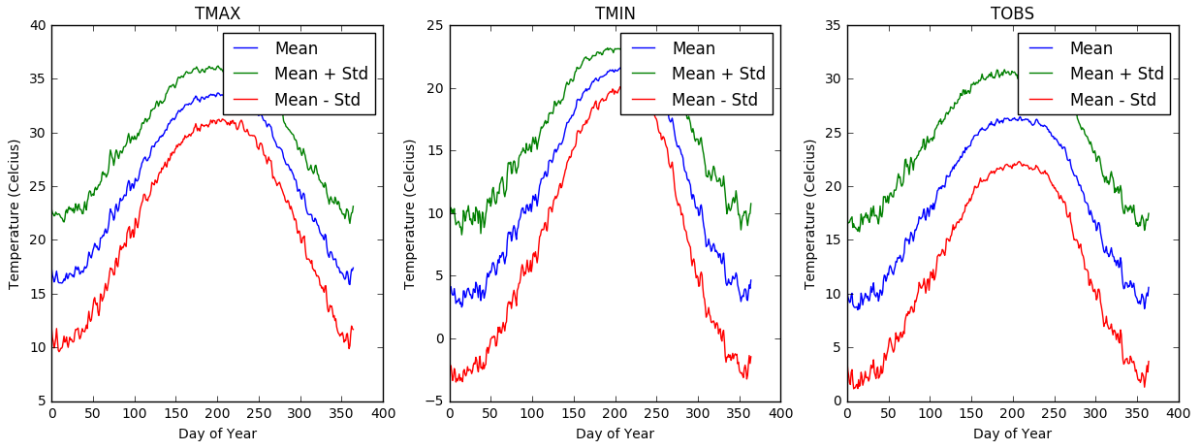


We see no major geographical features in the region of interest and use this location information to extract weather data from sources on the internet. This data is used in the data validation experiments. We note that the region experiences precipitation throughout the year at low intensity and rarely has snowfall. Temperatures vary from maximum of 85 Fahrenheit in the summer to a minimum of 35 Fahrenheit in the winter. In the following sections: (1) We validate our data using external sources (2) Perform some general analysis/observations from the data (3) Analyze the Snow Depth data (4) Analyze the temperature data (5) Analyze the precipitation data (6) Analyze patterns in how stations generate data

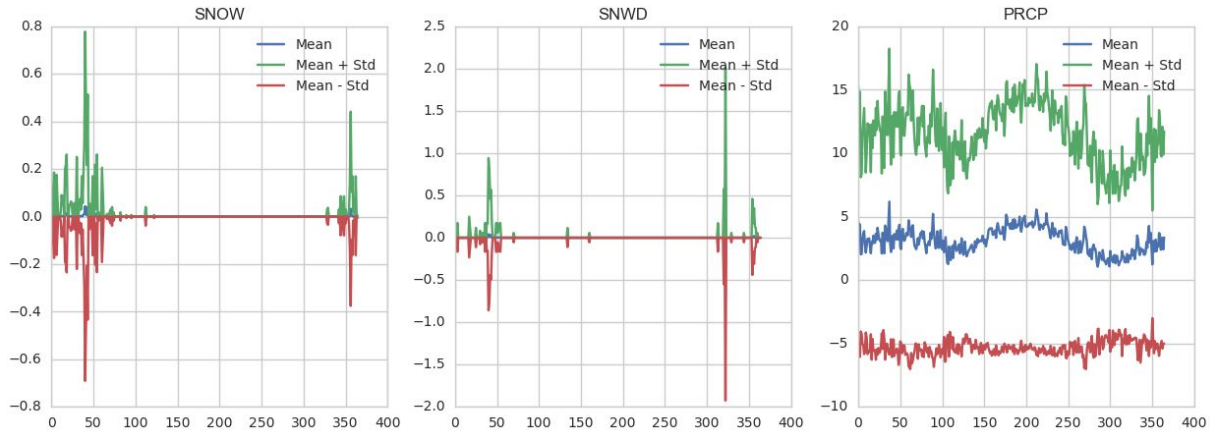
Data Validation Experiments

To perform initial sanity check experiments, we use the mean and standard deviation of the temperatures, snow depth and compare them to the data for the Georgia region from https://weatherspark.com/y/15598/Average-Weather-in-Atlanta-Georgia-United-States as well as literature about the weather in Georgia online.

The figure below shows the mean for the maximum, minimum and observed temperatures as a function of the day of the year. We see that these numbers agree with the data from weatherspark. We also see an increase in temperatures in the summer time in the middle of the year and a fall in temperatures in the winter periods.



Next we look at the snow depth, actual snowfall and the precipitation patterns on similar plots.

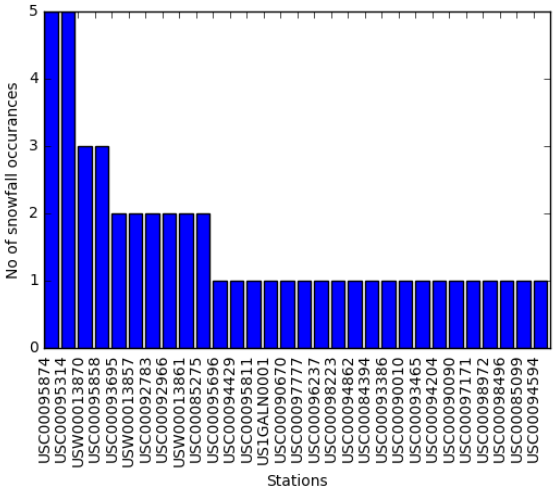


We see that the mean precipitation is around 4 throughout the year. In addition, the amount of precipitation sees a spike during June, July and August. This is again in line with the wet season in Georgia and the % chance of precipitation at the weatherspark site. We see that the mean snow and snow depth are very low throughout the year. This is again similar to average weather from the weatherspark site. In addition, we see that the standard deviation of snow depth and snow is very high compared to the mean. We will investigate this in the section on snowfall. The high standard deviation occurs in December, Jan and Feb i.e the winter season. The code for these experiments can be found in the 'Validation' notebook.

Analysis of Snow Depth

In this section, we first we talk about some general analysis on the snow depth data from this area. Then we perform mean and eigenvector analysis followed by an analysis of spatial vs temporal patterns in the data.

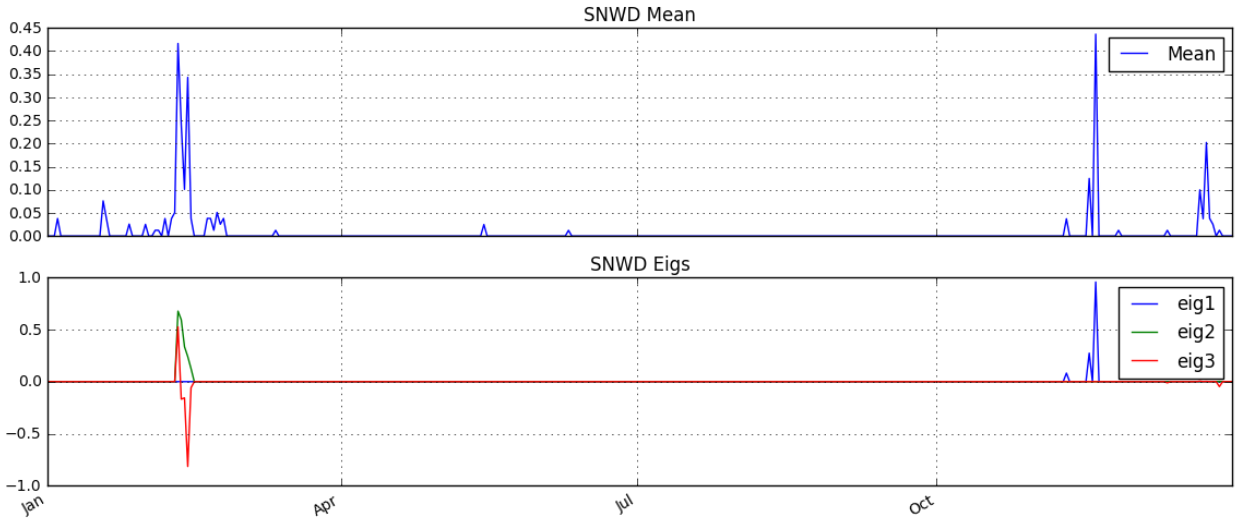
First we check for year-station combinations where the snow depth was non zero in at least on instance. We see that our of ~ 2000 combinations, just 48 have had some snowfall. This implies that snow is rare in this region and we further try to see if we can isolate snowfall to a few locations or a period in time. First we check the number of unique stations that have snowfall and we see that there are 30 stations. The figure to the right shows that number of instances of snow fall for these stations. From the figure we can see that although some areas seem to be more likely to receive snow, there is no significant concentration of the snow fall to just a few places.



The years where snow occured are [1958. 1960. 1961. 1967. 1968. 1973. 1977. 1978. 1979. 1980. 1985. 1988. 1989. 1992. 1993. 1996. 2000. 2002. 2009. 2010.]. This indicates that snowfall was fairly consistent with time and there was no significant pattern. We now proceed to perform the other two experiments which further validate some of these observations.

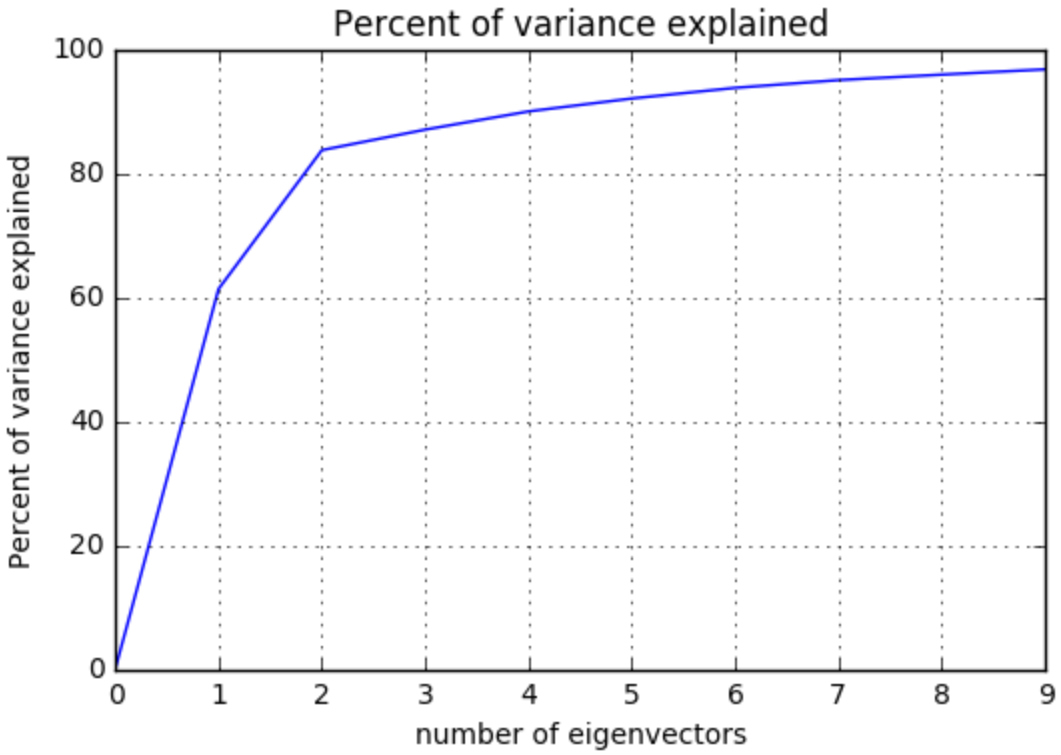
Mean and Eigenvalue Analysis: Temporal patterns

In this experiment, the eigenvalues and mean of snow depth as a function of day of year. The procedure is similar as that discussed in class.



The figure above shows that the mean of snow depth and the first three eigenvectors. The mean snow depth shows spikes in November, late December and February and March. This is essentially the winter season. The first eigenvector captures the snowfall late in the year while the second eigenvector captures the snowfall in and around March. The third eigenvector allows the possibility to capture the possibility of early snow. I.e. adding more and more of the third component, cancels out snow in late periods where the second component is non zero while adding more snow to earlier periods where the second component is non zero.

The plot below shows that cumulative variance explained by the eigenvectors. We can see that using three components, we can explain most of the variance present in snow depth.



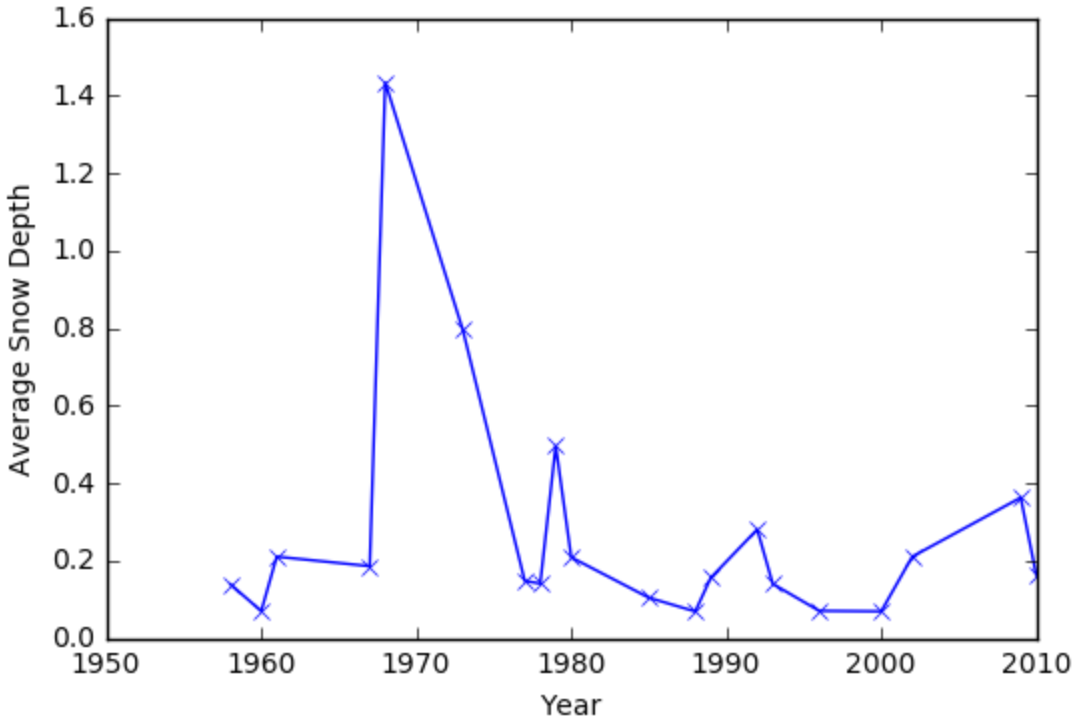
Spatial vs Temporal patterns

In this experiment, we try to determine if most of the variance in snow-depth is explained by variations across stations (Spatial) or across time (Temporal). To do this, we calculate the RMS overall and the reduced RMS after remove row wise means (Spatial pattern) or column wise means (Temporal pattern).

Total RMS = 184.528009373
RMS removing mean-by-station= 177.649674688
RMS removing mean-by-year = 76.1877263768

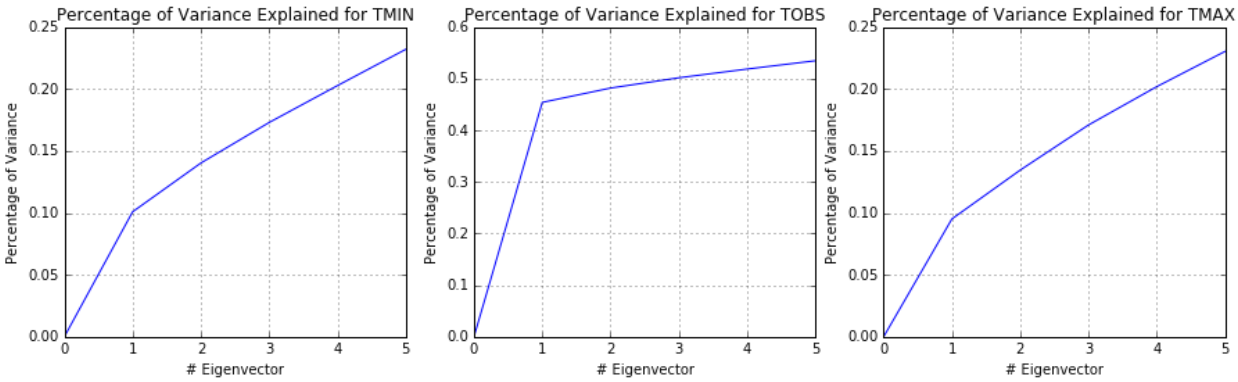
We see that the mean by year has the most effect and causes a huge reduction in RMS. To check and see if there are some patterns in yearly snow depth, we plot the average snow depth over years. The figure below shows this plot. We see that there is a large peak around 1970 and inspecting the actual values shows that the peak is actually at 1968 where the snow depth was 1.43. This contributes the largest amount to the RMS and explains why the per year means perform much better than per station means. However, on checking

weather reports, we see that this corresponds to a snow storm that was a one time event which has not recurred since on that scale. Thus this point is an outlier and since we have such a small sample size, it affects our results considerably. Thus the observation does not correspond to any pattern.

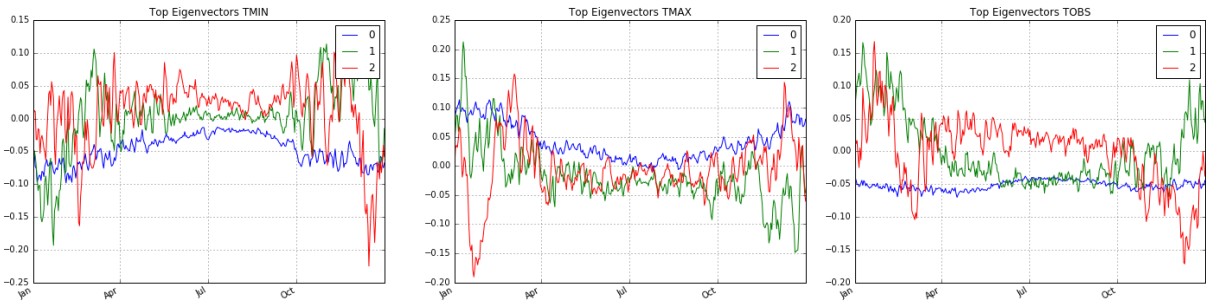


Analysis of Temperature Data

From the mean analysis performed earlier, we saw that the min, max and observed temperatures follow a pattern where all three show increases in the middle of the year i.e the summer months. Here we perform the PCA analysis of the three temperature measures. The variance explained as a function of the number of components is shown below.



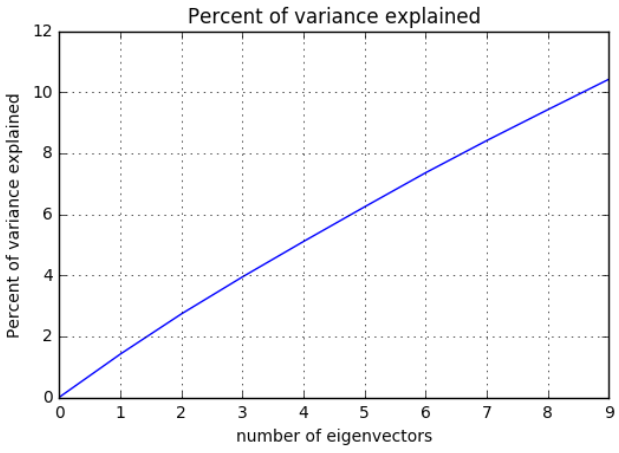
We see that for TOBS, the first 5 eigenvectors explain more than 50% of the variance. The first eigenvector in particular explains a significant portion of the variance. The TMIN and TMAX values are more noisy in comparison and only 20% of the variance is explained by the first 5 components. The figure below shows the first three eigenvectors.



The first eigenvector of TOBS shows peaks in both ends of the year while the mean remains fairly flat throughout. The first eigenvector explains away a lot of the variance in terms of change in temperatures between the summer (central part of the year) in relation to the winter (more extreme parts of the year). The degree of this change depends on the station and is a major source of variance in the data. The first component with spiked at both ends could help capture this variance among stations. The eigenvectors of TMIN and TMAX do have patterns albeit very noisy ones.

Analysis of Precipitation Data

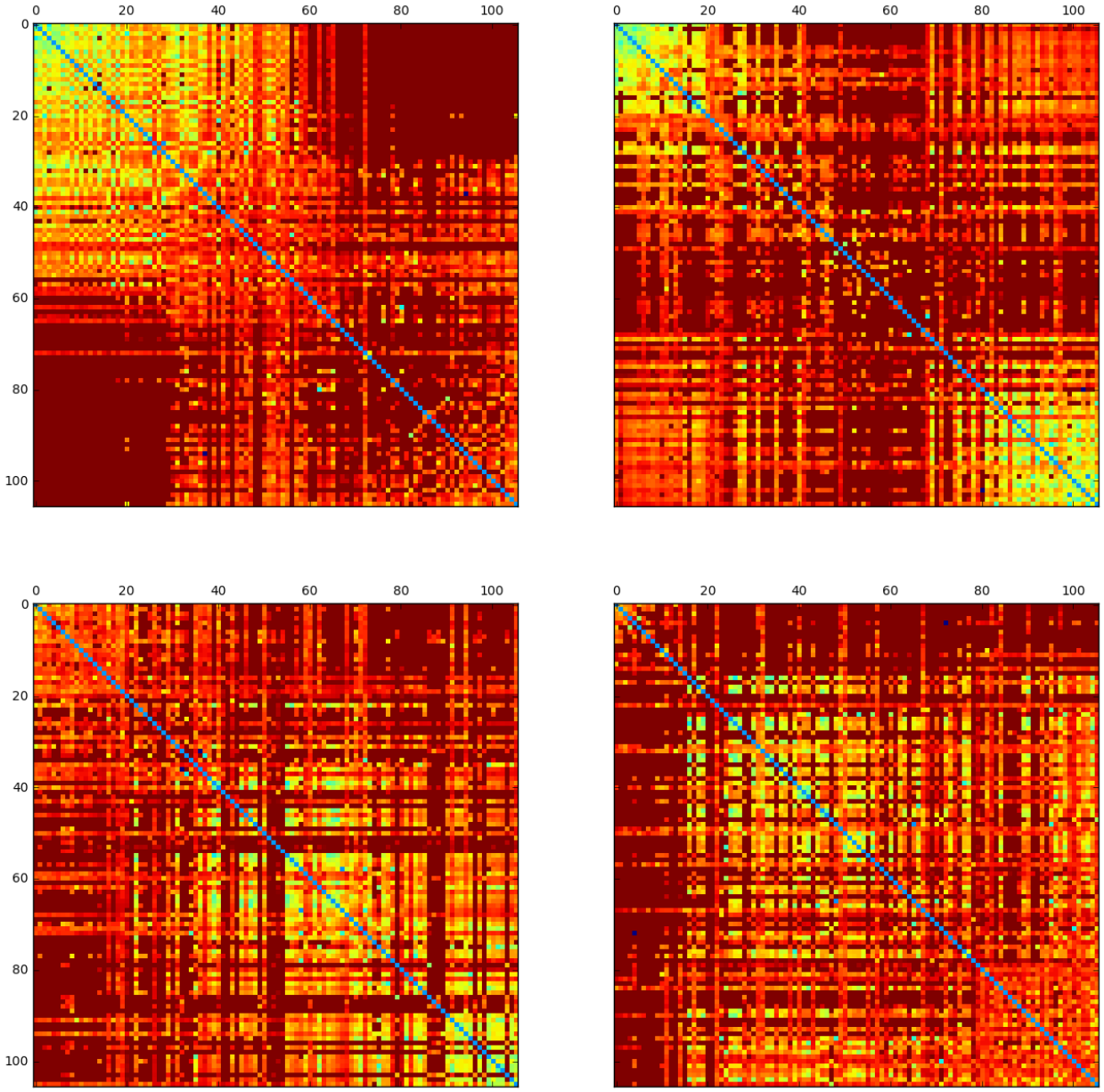
The figure below shows the mean and first three eigenvalues of the precipitation data over months of the year.



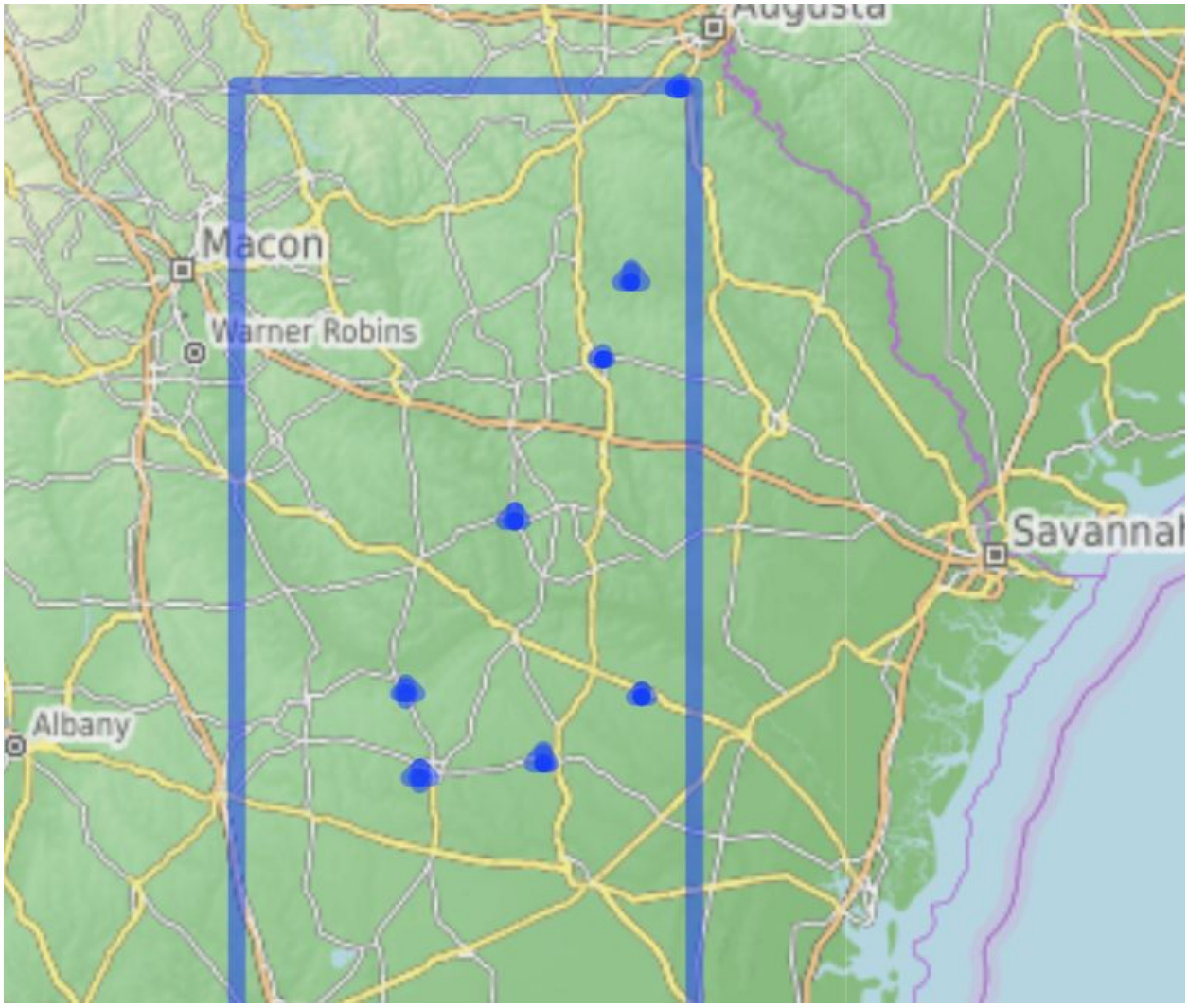
We do not see any significant patterns in the first three eigenvectors. We see that the mean shows a peak in the months of June, July and August. This is intuitive since this is the wet season in Georgia. The figure to the right shows the cum % variance explained. We see that the first 5 components explains less than 10% of the variance. This is again intuitive since there is a constant amount of precipitation throughout the year and there is no significant pattern to the precipitation data.

Analysis of correlation among precipitation occurrences:

We analyze the correlation of precipitation occurrences among stations. This uess notebook number 7 to perform the analysis. The correlation plot among stations is shown in the figure below:



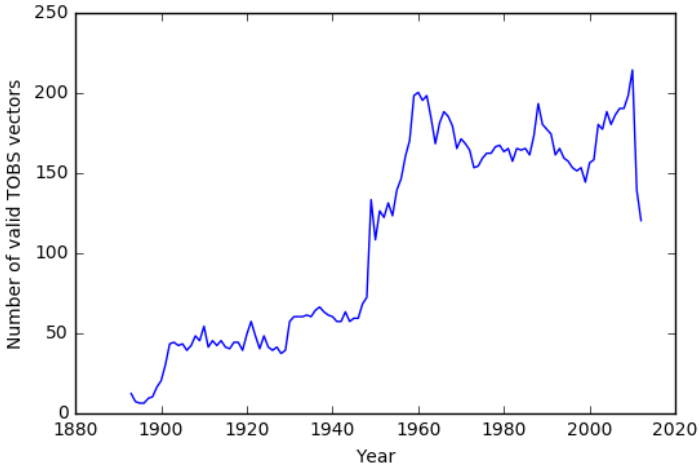
We see that there is block diagonal structure among stations 0-20 and 80 -100 in the first eigenvector. We explore further by extracting the station set with the most correlation and plotting the stations on the map:



We do not see any particular geographical features that can explain the correlation. However, precipitation patterns could be a function of prevailing wind, sea temperatures and might in addition depend on the type of precipitation the region receives etc. So we do not further attempt to reason about this as it would require significant domain knowledge.

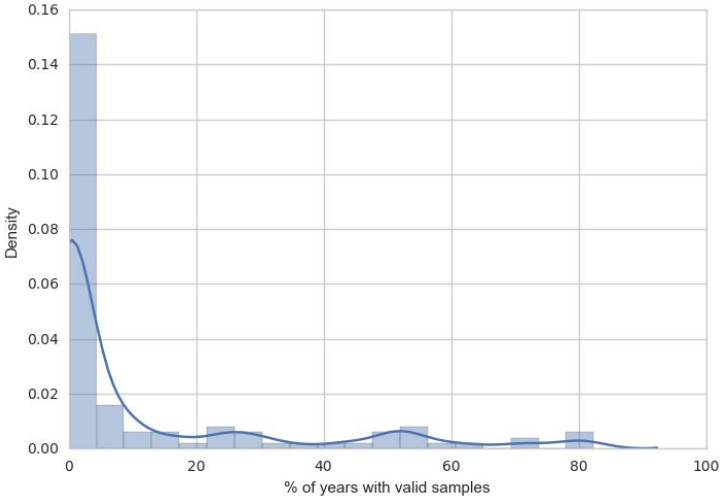
Station Sample Generation Patterns

The figure below shows the number of valid TOBS vectors we have as a function of the year. We see that there is big jump in the number of valid vectors we obtain at around 1950. Our hypothesis is that there are many stations introduced additionally around this time.

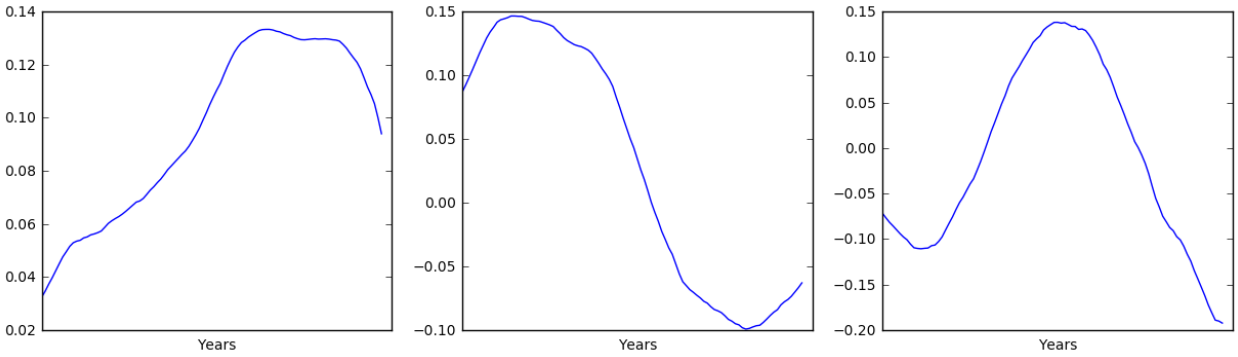


We measure the % of the year range that a station has a valid sample in terms of TOBS. We choose TOBS because we expect TOBS to have a valid sample when the station was present. This is in contrast to SNOW or PRCP etc which might have more variance.

We see that a large mode in % years close to 0%. Other modes are close to 20% and 50%. The hypothesis about 50% mode is that these are stations introduced post 1949 as per the pattern observed earlier. The higher percentage are observations for stations that were present throughout. We next analyze how the pattern of station presence looks using a PCA analysis to try validating our earlier hypothesis that there are many new stations introduced after 1949. To do this, we form a binary vector for each station with one entry per year in our range of interest. The entry indicates if we have a valid observation of TOBS in that year



from that station. This gives us a dataset of binary vectors one per station. A PCA analysis is performed on this data and we see that the first two eigenvectors explain more than 80% of the variance. The first three eigenvectors are shown in the plot below.



We see that the first eigenvector represents a station that occurs primarily post 1949 while the second eigenvector represents a station that occurs pre 1949 and does not occur post 1949. These two together explain most of the variance observed. Thus we conclude that 1949 was a year where a significant number of weather stations were added in the area of interest. The analysis here can be found in the 'Station Analysis' notebook along with some additional details about implementation as well some more plots exploring the station occurrence patterns.