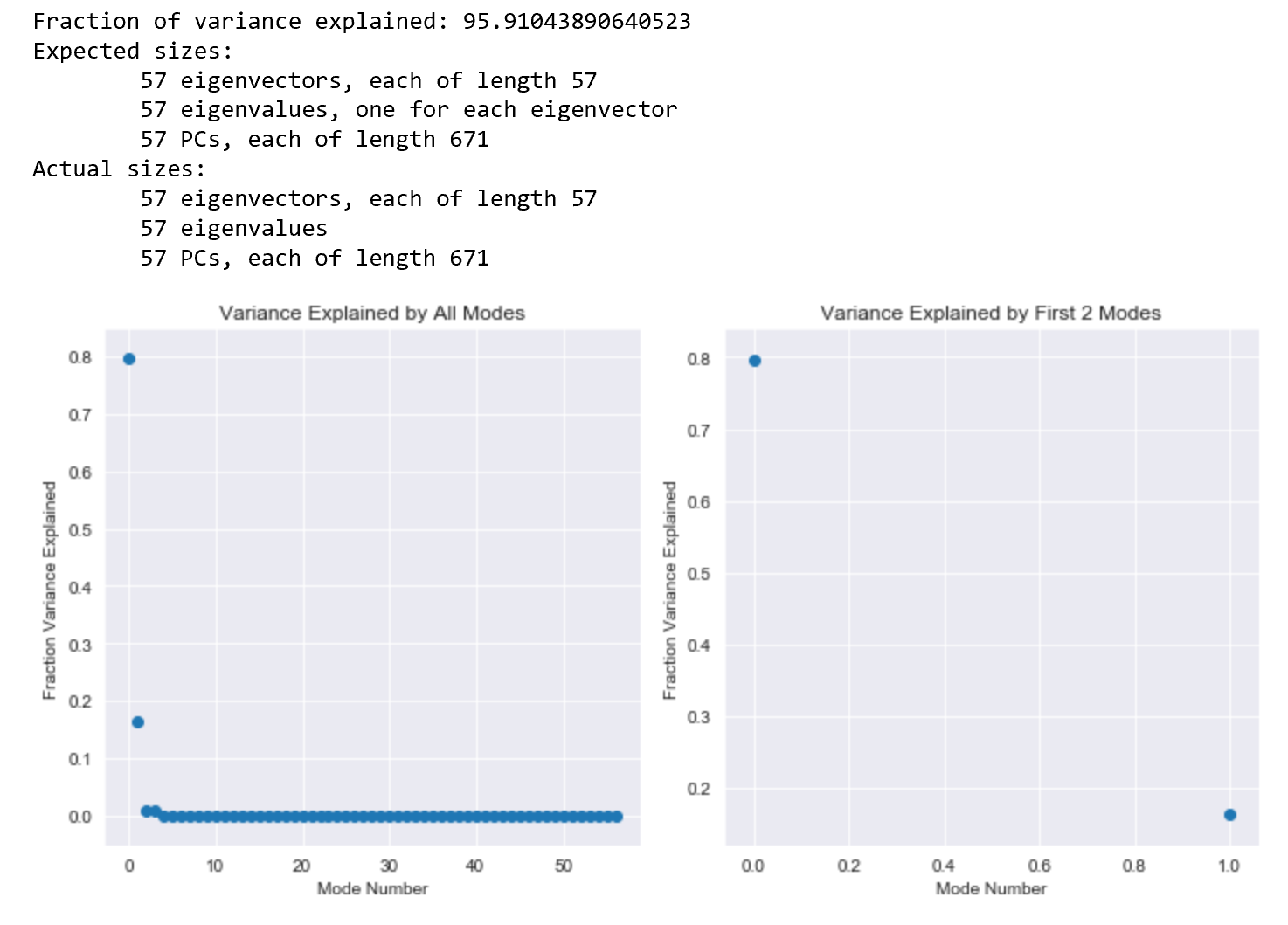
EOSC 510/410 Assignment 3

Problem 1:

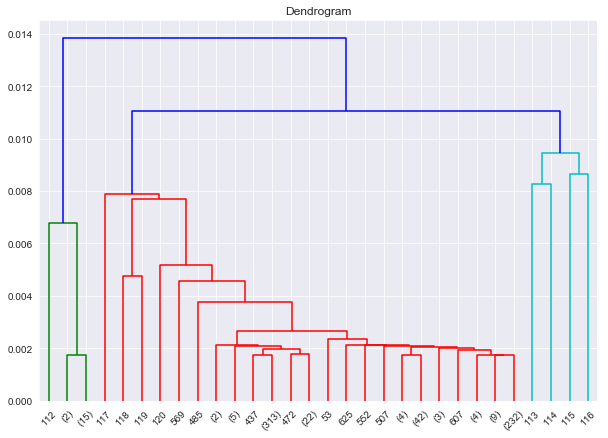
1. **Perform PCA on the data (the data has 57 variables and 672 observations in time) and decide how many modes to keep.**

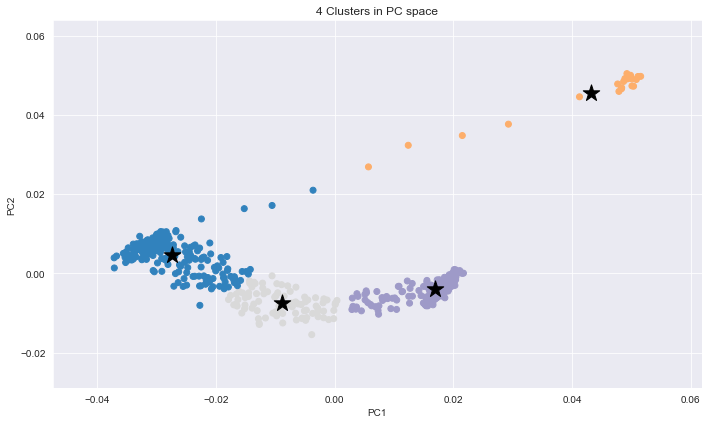
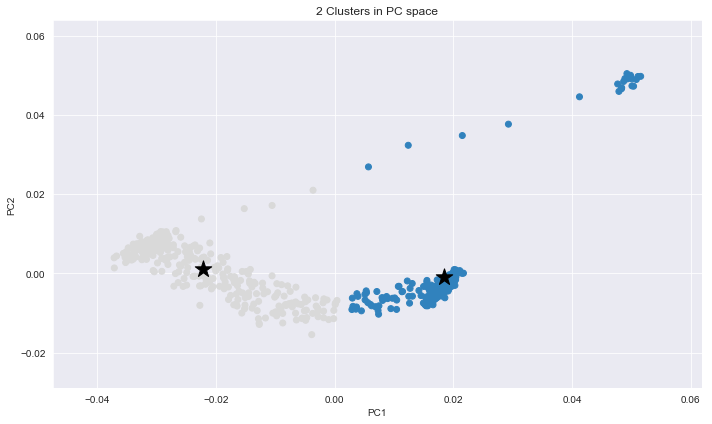
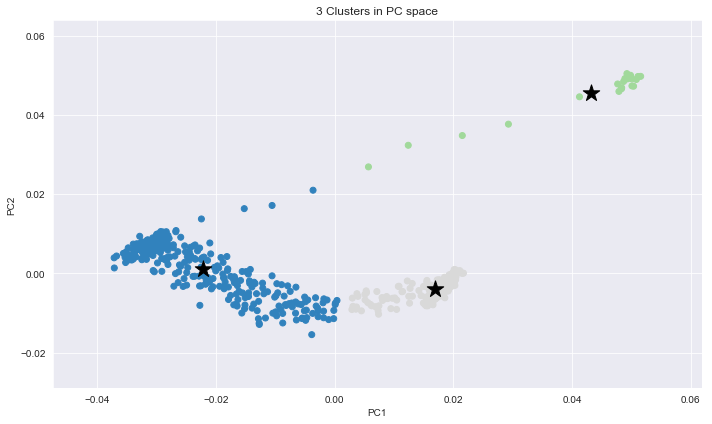
The output of PCA on this was the following:



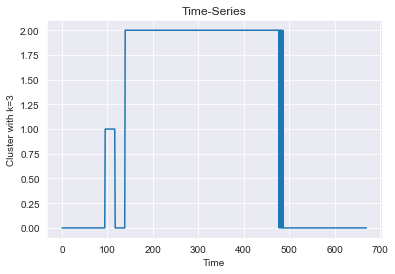
From figure 1 I decided to keep 2 modes since they described 95.9% of the variance.

1. **Perform hierarchical clustering with Ward's method on the data in the PC space of the modes you kept. Plot the dendrogram.**

****

1. **Chose three possible options for the optimal number of clusters (k) and plot the results (clustered data in PC space) for those options. [1 point for the correct choices of k, 1 point for the plots]** 

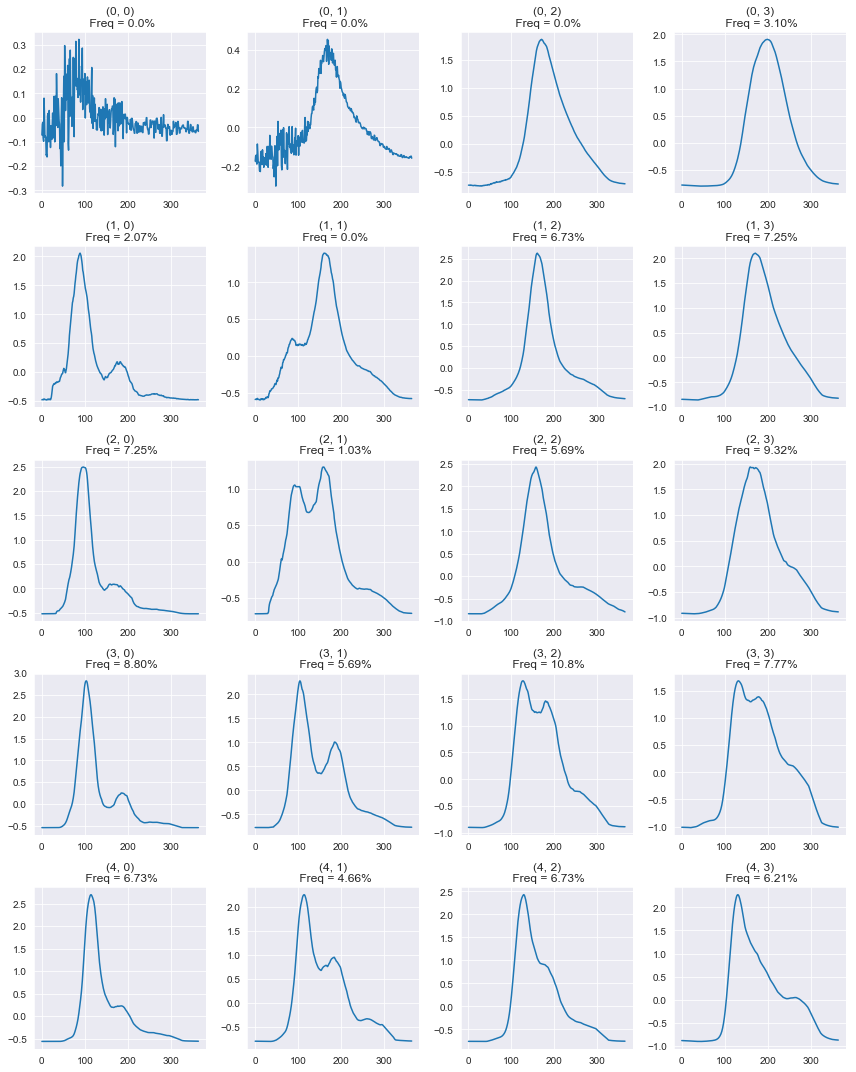
1. **For only one of the cluster options above (on choice of k): plot, on the same graph, the mean pattern (57 variables) of each cluster (using the reconstructed data according to the selected number of PC modes). Plot the time-series (672 points) of occurrences of these clusters. [1 point for the mean patterns plot, 1 point for the time-series plot]**

****

Problem 2:

1. **Perform clustering using a 3 x 2 SOM. Plot the 6 SOM patterns. Plot the locations of the stations, coloured according to the cluster to which they belong. What is the frequency of each cluster? [1 point for correct SOM patterns, 1 point for map of clusters, 1 point for correct frequencies]**

1. **Perform clustering a differently sized SOM, and plot the SOM patterns, locations of stations coloured by BMU, and frequency of each cluster as in a). Discuss what you think are two key differences between your results from a) and b). [1 point for plots, 2 points for discussion]**

****

Difference between the 2 SOMS:

1. Some patters have very low frequency in the 5 by 4 SOM (Such as (1,3),(2,2) and (3,3) at only 29%, 44% and 44% frequency)

2. The 3 by 2 SOM is often merging patterns so the (0,0) and (2,1) plots do not look much different. Less detain is revealed in this way. The 5 by 4 SOM however shows a bigger difference between the (0,0) and (4,3) plots. Some of the nodes however show more noise such as (0,0) or (0,1) since the patterns are fitting to more observations, and the BMUs try to match almost every observation.

1. **Calculate quantization error and topographic error for a range of SOM sizes (e.g.: 1x2, 2x2, 2x3, 3x3, 3x4, 4x4, 4x5, 5x5) and discuss what you find. In what circumstance is it more important to minimize quantization error, versus in what circumstance is it more important to minimize topographic error? [1 point for discussion of QE and TE with map size, 1 point for discussion on circumstances to minimize QE/TE]**

| **SOMs** | **QEs** | **TEs** |
| --- | --- | --- |
| 1x2 | 0.149 | 0.000 |
| 2x2 | 0.092 | 0.000 |
| 2x3 | 0.064 | 0.005 |
| 3x3 | 0.048 | 0.026 |
| 3x4 | 0.040 | 0.098 |
| 4x4 | 0.033 | 0.041 |
| 4x5 | 0.029 | 0.078 |
| 5x5 | 0.024 | 0.093 |

Larger maps have smaller QE, since there are more nodes which can represent the observations. The TE however increases with size of the maps. The QEs need to be minimized if we want to look at changes over time. In environmental or geoscientific data, we might want to look at timeseries data where we want to detect patterns over time. Since QE tells us how well the BMUs represent the observations, it is important to minimize the QE when looking for small and distinct changes over time. This is because in order to detect change, it is important that the BMU is allocated correctly. Since the TE determines how well does the map place similar patterns beside each other, minimizing TE might be more important while looking for spatial patterns or clusters in data. In general larger maps tend to have larger TE, since nearby nodes have more similar patterns.

**d) Calculate quantization error and topographic error for pairs of SOMs which have the same number of nodes but different map sizes and discuss what you find (e.g.: are QE and TE the same for a 2x3, 3x2 map, and 1x6 map? A 3x4 and 4x3 map? A 4x5 and 5x4 map? A 1x2 and 2x1 map?). [1 point for identifying if QE/TE are the same for maps with the same number of nodes and different shape, 1 point for discussion]**

| **SOMs** | **QEs** | **TEs** |
| --- | --- | --- |
| 1x2 | 0.158 | 0.000 |
| 2x1 | 0.162 | 0.000 |
| 2x3 | 0.066 | 0.000 |
| 3x2 | 0.068 | 0.000 |
| 1x6 | 0.069 | 0.021 |
| 3x4 | 0.040 | 0.057 |
| 4x3 | 0.042 | 0.161 |
| 4x5 | 0.026 | 0.109 |
| 5x4 | 0.027 | 0.000 |