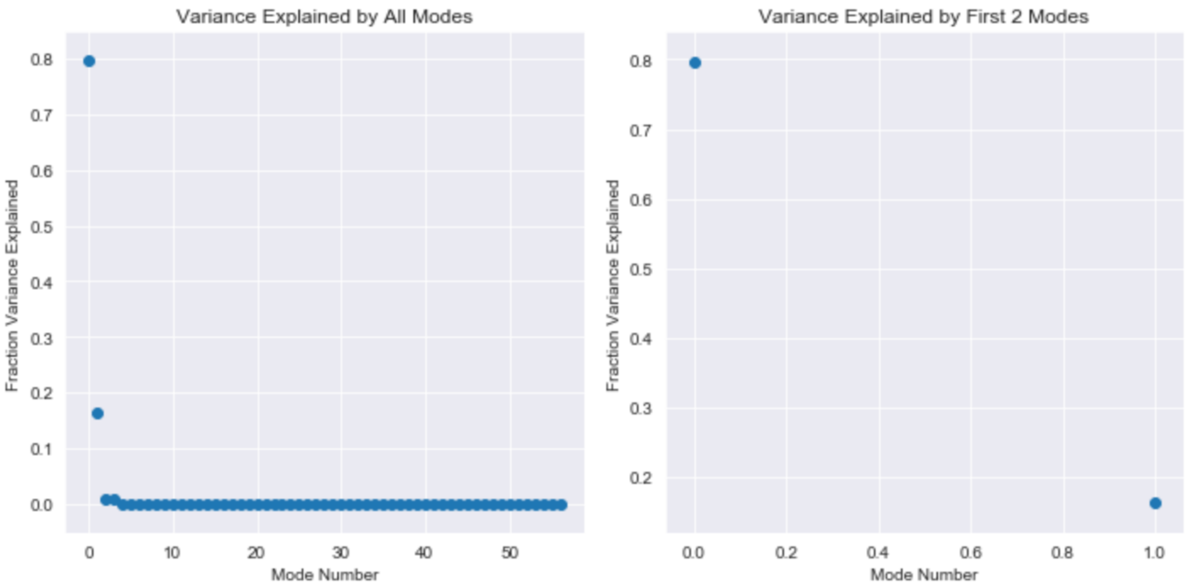
EOSC 510/410 Assignment 3

Problem 1:

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

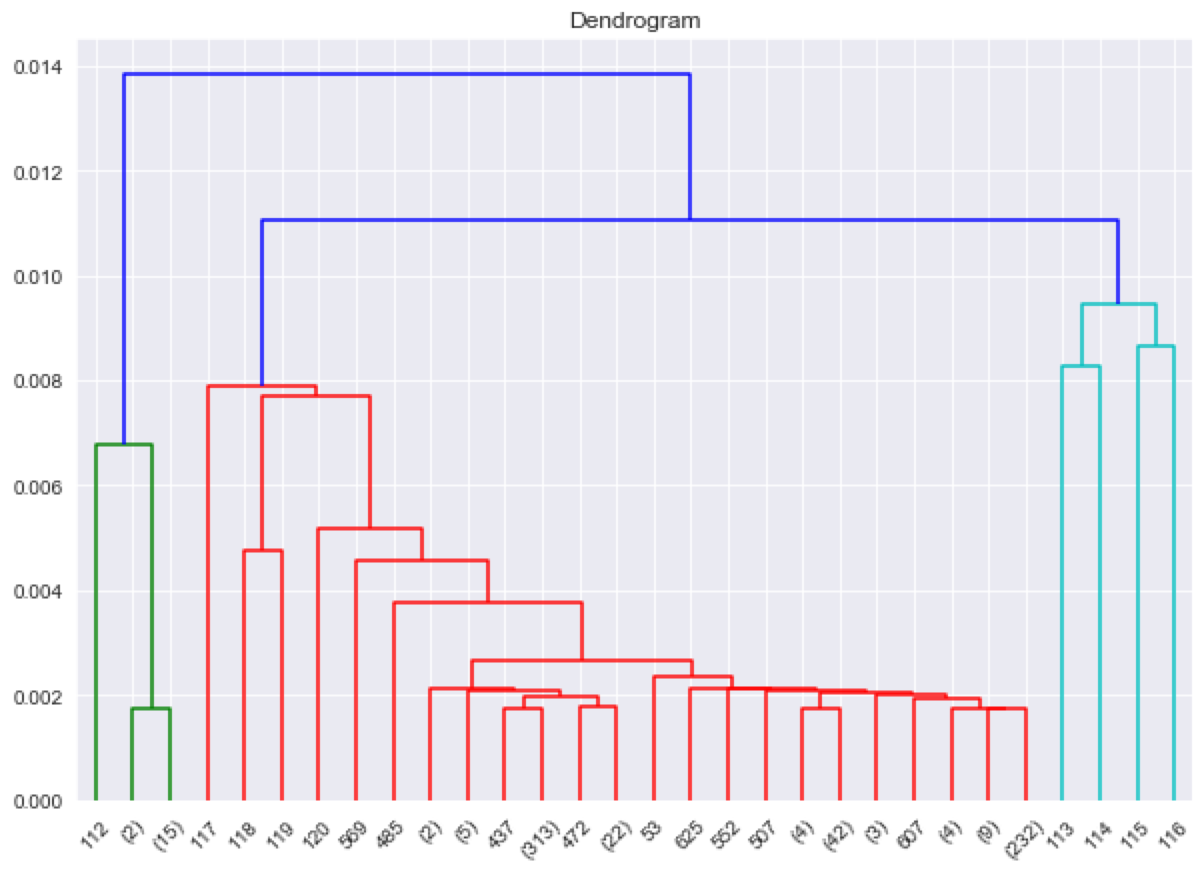
The output of PCA on this was the following:

```
Fraction of variance explained: 95.91043890640523
Expected sizes:
        57 eigenvectors, each of length 57
        57 eigenvalues, one for each eigenvector
        57 PCs, each of length 671
Actual sizes:
        57 eigenvectors, each of length 57
        57 eigenvalues
        57 PCs, each of length 671
```

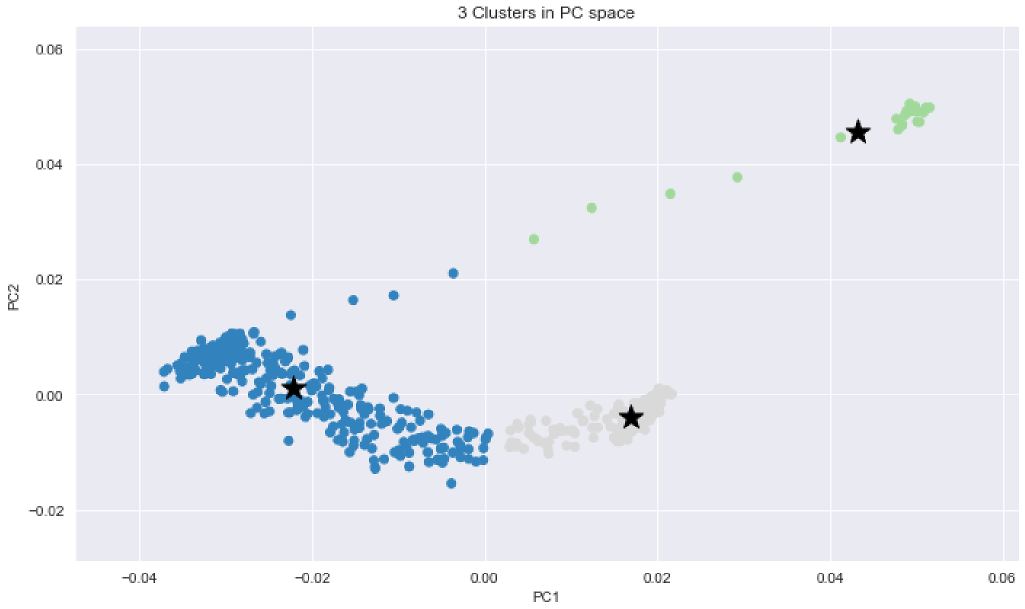


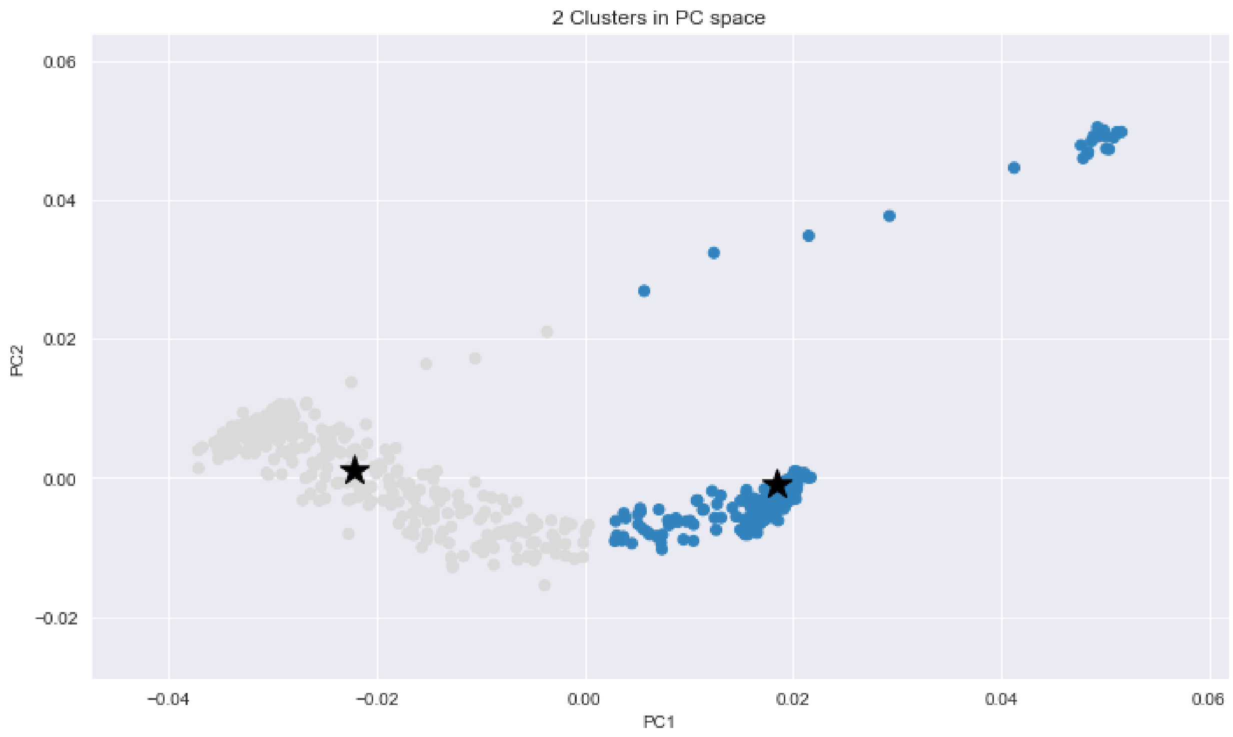
From figure 1 I decided to keep 2 modes since they described 95.9% of the variance.

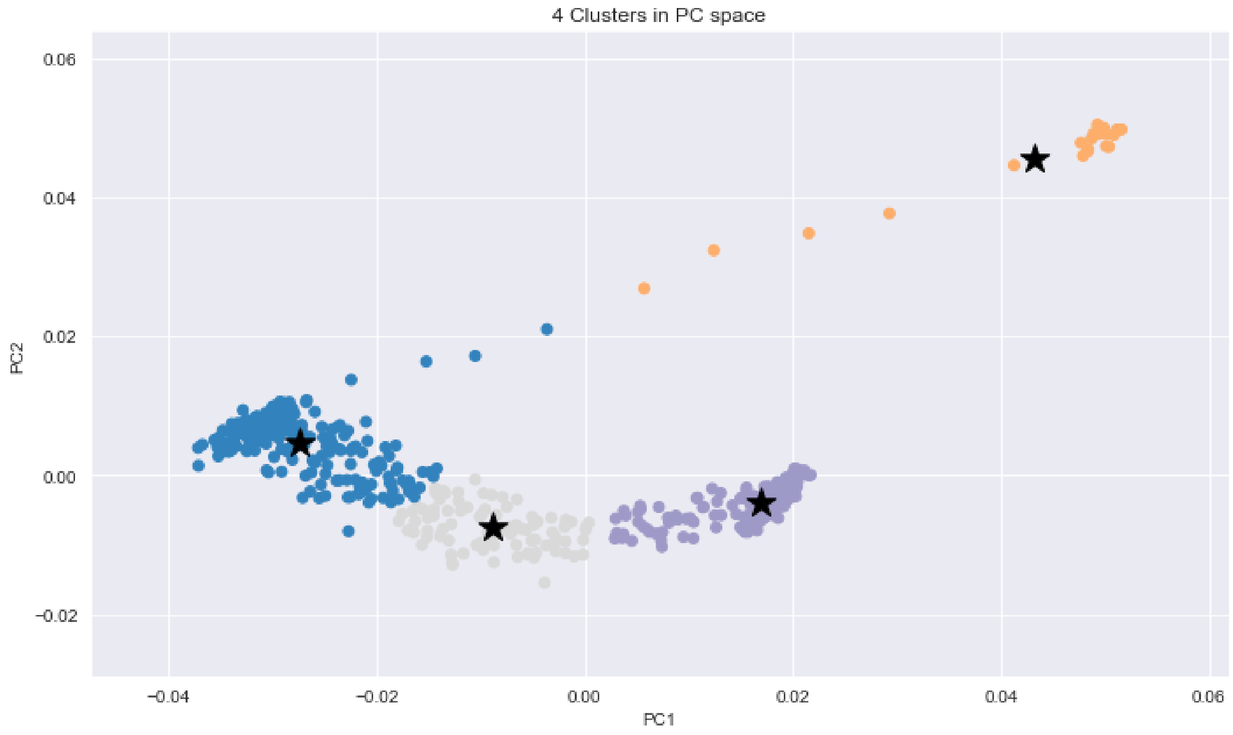
b) **Perform hierarchical clustering with Ward's method on the data in the PC space of the modes you kept. Plot the dendrogram.**



c) **Chose three possible options for the optimal number of clusters (k) and plot the results (clustered data in PC space) for those options.**

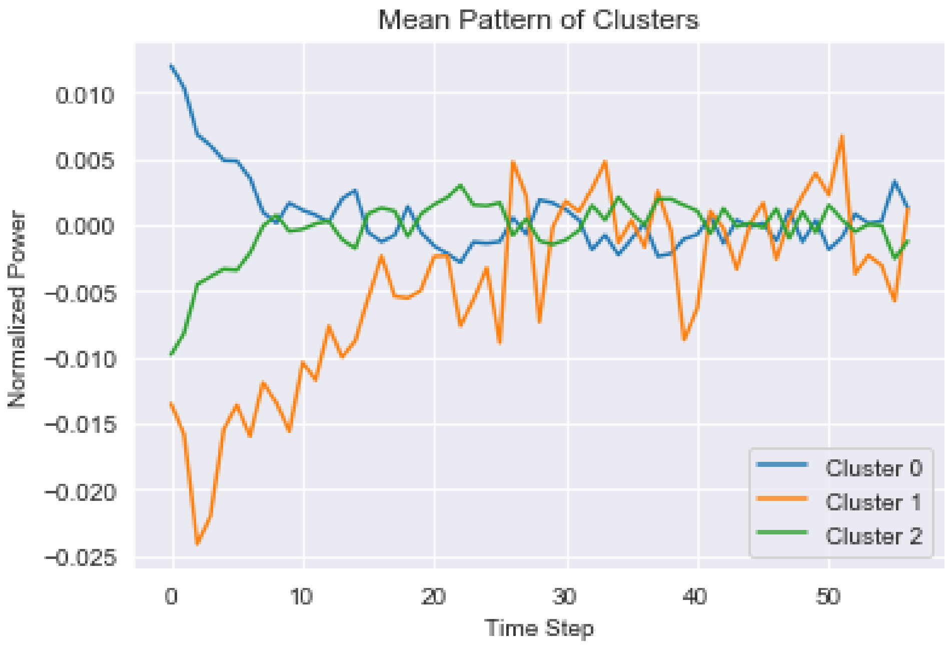


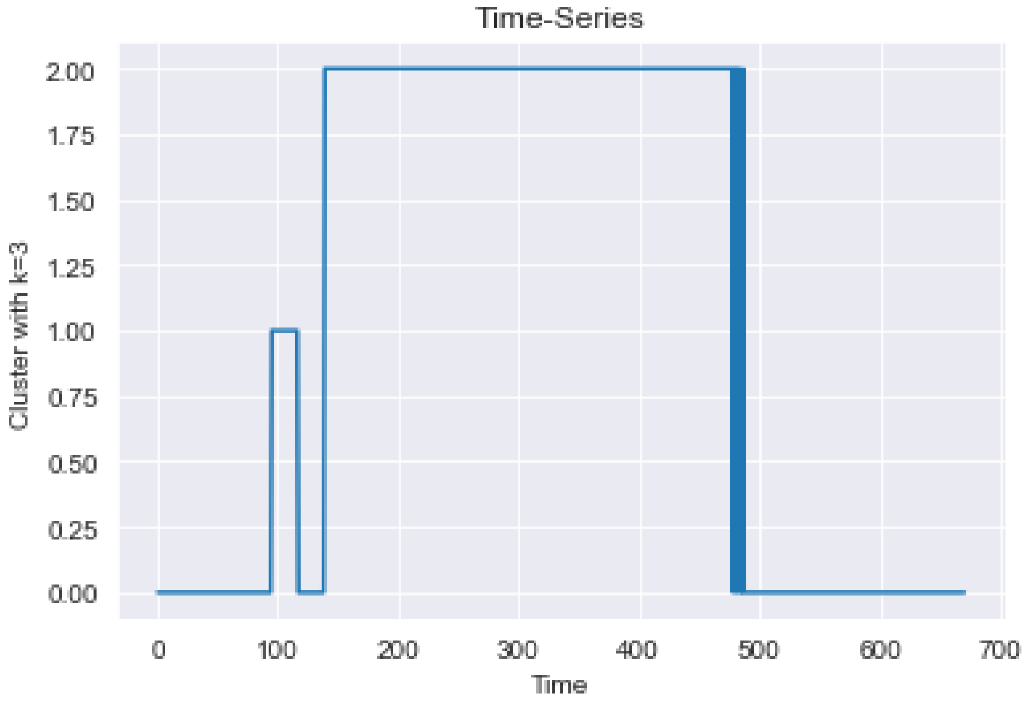




d) For only one of the cluster options above (on choice of k): plot, on the same graph, the mean pattern (57 variables) of each cluster (using the reconstructed data according to the selected number of PC modes). Plot the time-series (672 points) of

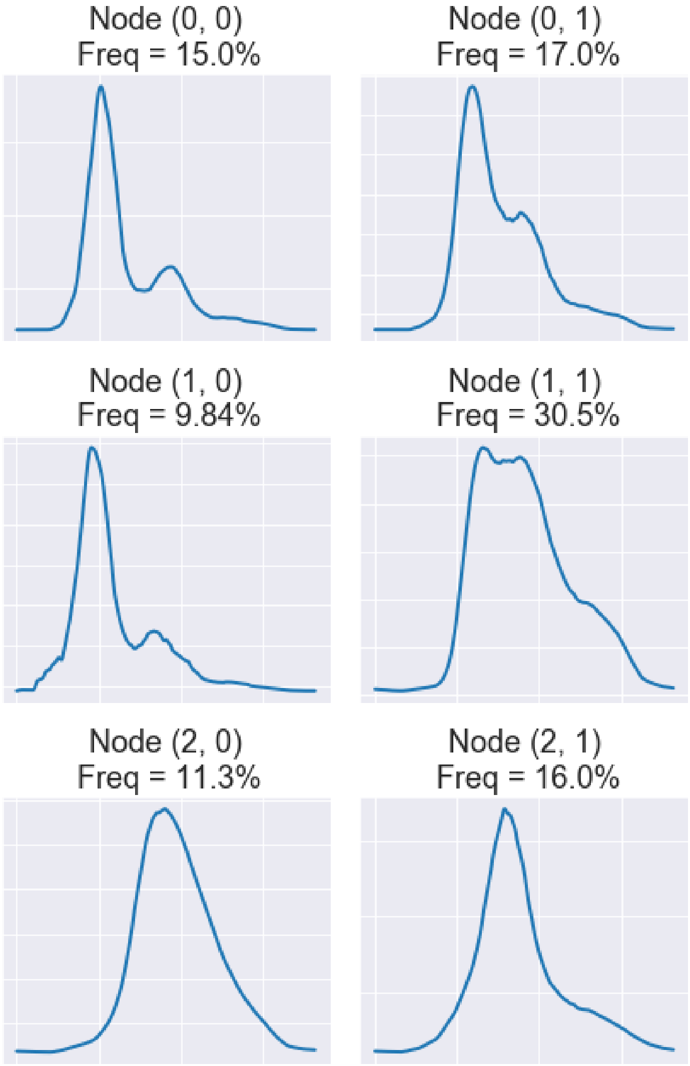
occurrences of these clusters. [1 point for the mean patterns plot, 1 point for the time-series plot]

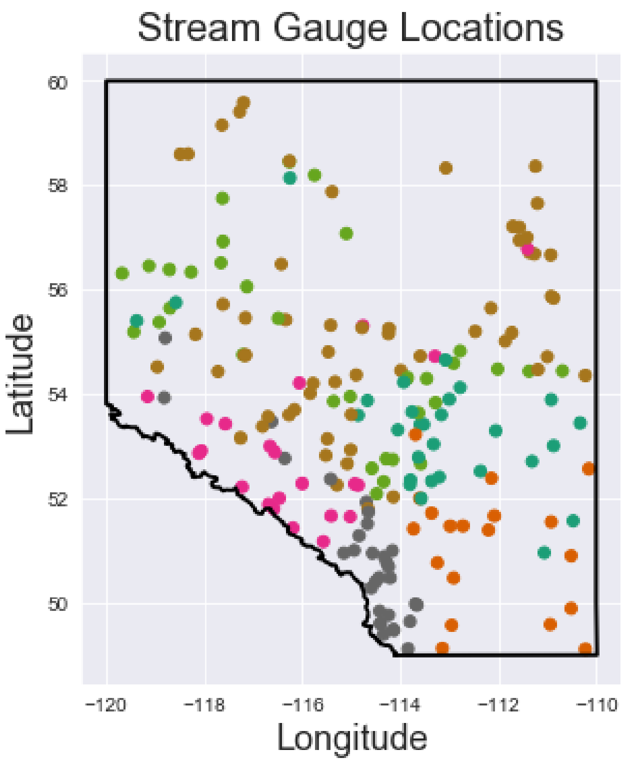




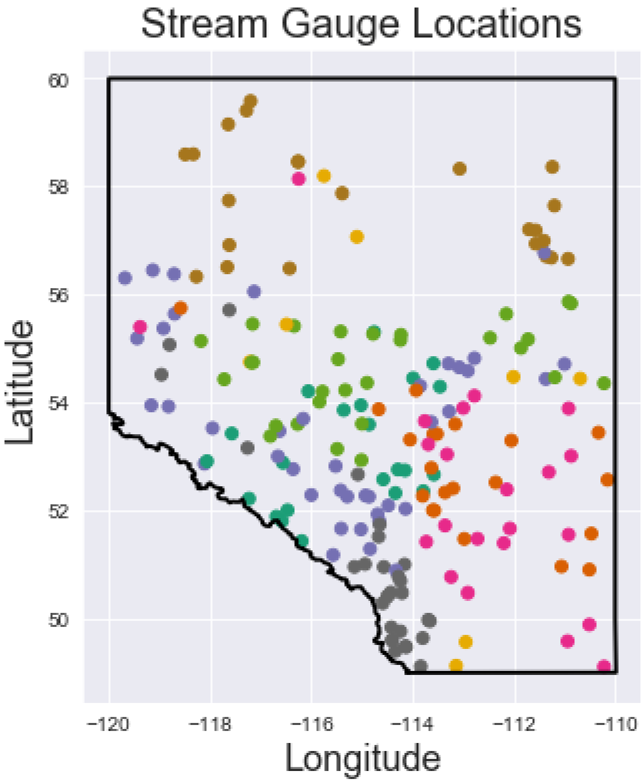
Problem 2:

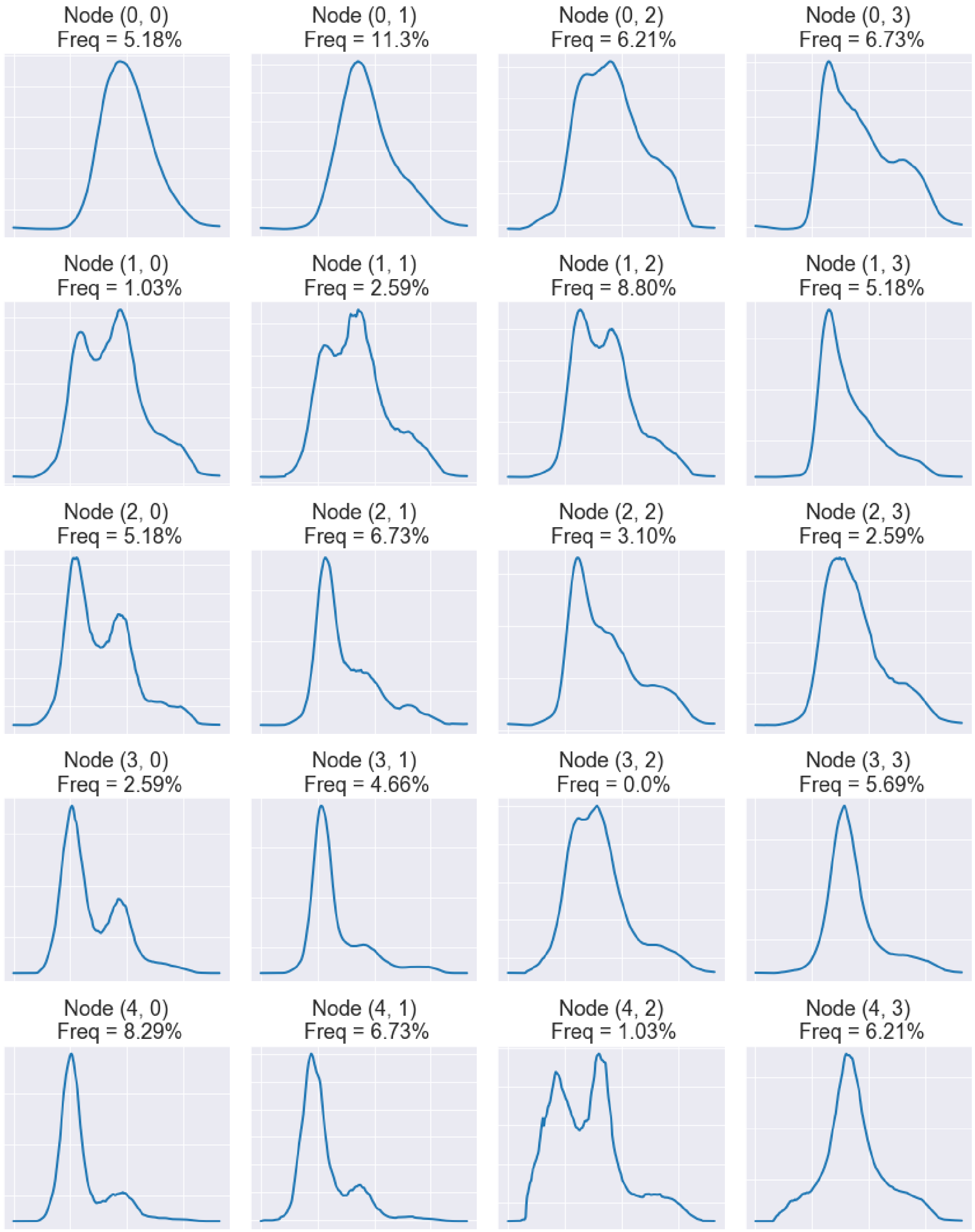
a) **Perform clustering using a 3 x 2 SOM. Plot the 6 SOM patterns. Plot the locations of the stations, coloured according to the cluster to which they belong. What is the frequency of each cluster?**





b) 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).


Stream Gauge Locations
Latitude
60
58
56
54
52
50
−120
−118
−116
−114
−112
−110
Longitude



Node (0, 0)
Freq = 5.18%

Node (0, 1)
Freq = 11.3%

Node (0, 2)
Freq = 6.21%

Node (0, 3)
Freq = 6.73%

Node (1, 0)
Freq = 1.03%

Node (1, 1)
Freq = 2.59%

Node (1, 2)
Freq = 8.80%

Node (1, 3)
Freq = 5.18%

Node (2, 0)
Freq = 5.18%

Node (2, 1)
Freq = 6.73%

Node (2, 2)
Freq = 3.10%

Node (2, 3)
Freq = 2.59%

Node (3, 0)
Freq = 2.59%

Node (3, 1)
Freq = 4.66%

Node (3, 2)
Freq = 0.0%

Node (3, 3)
Freq = 5.69%

Node (4, 0)
Freq = 8.29%

Node (4, 1)
Freq = 6.73%

Node (4, 2)
Freq = 1.03%

Node (4, 3)
Freq = 6.21%

Difference between the 2 SOMS:

1. Some patterns have very low frequency in the 5 by 4 SOM (Such as (3,2) ,(1,0), and (4,2) at only 0.0%, 1.03% and 1.03% frequency). This is because as more nodes are created, sometimes not all nodes have enough observations.

2. The 3 by 2 SOM is often merging patterns so the (0,0) and (2,1) plots do not look much different. Less detail 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 in the 5 by 4 since the patterns are fitting to more observations, and the BMUs try to match almost every observation.

c) 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. In general, one needs to minimize TE if they want similar patters to be placed closed to one another spatially (perhaps by using a hexagonal map), and minimize QE to ensure the BMUs fit more 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	0.161	0.000
2x1	0.154	0.000
2x3	0.071	0.000
3x2	0.065	0.000
1x6	0.067	0.181
3x4	0.041	0.083
4x3	0.039	0.098
4x5	0.026	0.026
5x4	0.028	0.083

The QE remains approximately the same for maps of the same total nodes. For instance a 2x3, 3x2 and 1x6 map all have 6 nodes. They all have approximately the same QE. If rounded to 2 decimal places, all the maps shown in the table above that have the same total nodes, have the same QE. This is because in either configuration the number of BMUs depends on the number of nodes, which are unchanged. These BMUs then define the same observations in either case.

However, the TE differs between maps of different shapes. As anticipated the 1x6 map has the largest TE because the patterns are placed in a straight line, and therefor even if 1 and 3 had some resemblance, the are not placed closed to one another. Additionally as seen in the 4x3 and

3x4 case, maps that were “longer” than “wider” also have a bigger TE. I would assume this is because if similar patterns are placed left to right in a row and similar rows are stacked, then longer rows (wider maps) have more similar patterns placed beside one another. This would minimize their TE. Similarly, longer maps (4x3 or 5x4) will have a larger TE than their wider counterparts (3x4 or 4x5 respectively).