**Clustering**

Cluster analysis is used to determine whether groups exist in the data. Once the existence of groups is established, we then want to classify observations to groups. Clustering is an unsupervised method in which no structure is assumed, but instead we work to discover a structure by analysing the data provided.

We have looked at two different methods of clustering:

* Hierarchical Clustering
* Iterative Clustering (*k*-Means)

**Hierarchical Clustering:**

This method involves building a tree, where similar objects are joined low down and less similar objects are joined higher up. In this way, observations are placed in groups in accordance with similarity. The dataset we are investigating has 16 variables, all with differing units of measurement, therefore it is clear that the data needs to be modified before building our dissimilarity matrix.

We first removed the date column from the data as we required only numerical values for our analysis. The second column of the data contained a binary variable detailing whether a diagnosis was made on that day. In our initial analysis we removed that variable also an focused on the 14 variables relating to the weather conditions.

We then standardised the data to remove any potential sway caused by a variable with particularly large variance. Having calculated dissimilarity matrices using a number of proposed methods such as Euclidean, Manhattan and Maximum distance, we began to compose our dendrograms. This process was repeated using three proposed linkage methods: Single, Average and Complete. For each method we calculated a recommended cut-off height using the formula:

However regardless of the method the number of clusters was very high; ranging from 40 for Maximum Distance and Single Linkage, to 78 for Manhattan Distance and Complete Linkage. This result was perhaps misleading as the degree of chaining associated with our “best” outcome was quite severe.

Some further research lead us to the “cluster” library which contains a function, Daisy, capable of building dissimilarity matrices with variables of different type. We ran our analysis again using the Daisy function and including the diagnosis variable however the outcome was much the same with number of clusters ranging from 53 to 81.

We concluded that due to the wide variety of factors contributing to the weather, hierarchical clustering may not be able to provide a succinct number of clusters.

**Iterative Clustering:**

K-means is an iterative clustering method that assigns objects into k groups where the observations within a group are very similar and the groups are very different from each other. At each stage in the algorithm data points are assigned to a fixed number of k clusters. As the dataset is 16-dimensional, we could not select k by observation. Therefore we ran a loop for k equals 1 to 10 to calculate the sum of squared distances of each object to its cluster centroid. We then plotted k against SS and the results can be seen in Appendix X below.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2017).

cluster: Cluster Analysis Basics and Extensions. R package version 2.0.6.

