Cluster Analysis

This cluster analysis challenge involves studying how penguins impact the composition of soils and surface waters on King George Island (Maritime Antarctic Admiralty Bay). The goal is to identify clusters that reveal areas where penguins have influenced the environment at similar rates.

December 2021

Table of Contents

Abstract

This analysis attempts to group measured nutrient data into clusters to determine the areas that represent the common impact of penguins. The analysis was conducted using Microsoft Excel and R Studio, employing several distance methods combined with various clustering algorithms. These methods included Euclidean, Manhattan, Maximum, Minkowski, and Canberra distances, each paired with clustering methods such as Ward.D, Ward.D2, Single, Complete, Average, McQuitty, Median, and Centroid. A total of 40 dendrogram plots were generated from these combinations.

The generated dendrograms were analyzed, and a preferred plot was selected based on the dominant groupings observed. After further consideration and verification using Silhouette and Elbow plots, the data points were grouped into three clusters, which correspond to the number of penguin species and help indicate the areas with similar penguin impact. Overall, penguins impacted points 5 and 8 at similar rates, while point 15 was affected at a significantly different rate from the other points in the study area.

1.0 Introduction

This analyses made use of dataset 2, which is composed of measurement results of cations and nutrients in the terrestrial ecosystems of the Antarctic sea, that were impacted by the three penguins species. This analyses essentially seeks to find the areas (represented by the various measurement points) where the penguins commonly toll within the Antarctic sea. In order to achieve this, the cluster analyses were conducted with the aid of Microsoft Excel and R studio. The analyses began with careful observation of the of the data provided, this was necessary to ensure that there are no missing data within the data provided.

As shown in Table 1, Sodium (Na) appears to generally have highest percentages out of all the recorded nutrients, while Iron (Fe) has the lowest representation. Therefore if the analyses is conducted without having due regard for the difference in substances' concentrations, all the distances calculated will probably be more representative of the substances with higher percentages, at the detriment of those with lower percentages - That is, the distance analyses will be bias towards the concentrations with higher values, and this will render the result of the analyses inaccurate. Therefore, in order to avoid this, the data was normalised - each of the concentrations recorded were divided by the mean values of the sum total of that particular concentration as shown in Table 2. This was done to bring all the measured concentrations into somewhat equal scale, to ensure that the data are more comparable, and to eliminate bias in the calculation of distances and the consequent results of clustering.

Table 1: Showing the Raw Data Provided

	Na	K	Ca	Mg	Fe	Mn	Sr
1	27.01	1.57	21.03	6.51	0.17	0.02	0.92
2	27.68	1.30	8.03	4.36	0.44	0.03	0.96
3	28.87	8.35	18.46	13.17	0.30	0.31	1.13
4	14.52	1.62	11.13	1.63	0.28	0.03	0.99
5	20.05	17.55	11.58	51.86	0.32	0.89	1.12
6	29.93	10.69	16.56	11.32	0.25	0.25	1.07
7	24.77	0.70	9.20	5.44	0.23	0.02	0.93
8	48.63	37.01	33.16	19.99	0.27	0.67	1.38
9	25.00	4.57	1.31	1.32	0.28	0.02	0.66
10	31.40	4.55	16.58	9.31	0.68	0.13	0.85
11	34.28	6.95	14.82	9.85	1.17	0.30	1.21
12	35.68	13.64	23.51	13.90	0.79	0.49	1.32
13	24.08	1.62	6.69	5.37	0.68	0.04	0.92
14	14.18	1.04	0.63	0.99	0.49	0.04	0.88
15	26.85	2.90	5.94	4.62	5.67	0.16	0.90
16	70.59	9.49	4.41	3.75	2.00	0.17	0.86
17	22.48	0.45	6.17	2.55	0.26	0.02	0.28
18	127.57	7.52	5.13	10.27	0.58	0.03	0.45
19	20.12	1.29	3.52	2.52	0.59	0.03	0.45
20	17.31	1.20	2.87	1.96	0.45	0.02	0.61
21	23.92	3.01	6.72	3.53	0.29	0.06	0.78
22	56.51	20.01	9.86	8.16	0.30	0.06	0.88
Mean	**34.15**	**7.14**	**10.79**	**8.74**	**0.75**	**0.17**	**0.89**

Source: Environmental Monitoring Project Material by Loga Małgorzata 2021

Table 2: Showing the Scaled Data Matrix

	Na	K	Ca	Mg	Fe	Mn	Sr
1	0.79	0.22	1.95	0.74	0.23	0.13	1.03
2	0.81	0.18	0.74	0.50	0.59	0.15	1.08
3	0.85	1.17	1.71	1.51	0.39	1.79	1.28
4	0.42	0.23	1.03	0.19	0.37	0.17	1.12
5	0.59	2.46	1.07	5.93	0.42	5.20	1.26
6	0.88	1.50	1.53	1.29	0.33	1.45	1.20
7	0.73	0.10	0.85	0.62	0.31	0.10	1.05
8	1.42	5.19	3.07	2.29	0.35	3.90	1.56
9	0.73	0.64	0.12	0.15	0.38	0.13	0.74
10	0.92	0.64	1.54	1.06	0.91	0.73	0.95
11	1.00	0.97	1.37	1.13	1.56	1.78	1.37
12	1.04	1.91	2.18	1.59	1.05	2.88	1.49
13	0.71	0.23	0.62	0.61	0.90	0.26	1.03
14	0.42	0.15	0.06	0.11	0.65	0.22	0.99
15	0.79	0.41	0.55	0.53	7.57	0.93	1.01
16	2.07	1.33	0.41	0.43	2.67	0.96	0.97
17	0.66	0.06	0.57	0.29	0.35	0.10	0.31
18	3.73	1.05	0.48	1.17	0.77	0.15	0.50
19	0.59	0.18	0.33	0.29	0.79	0.16	0.51
20	0.51	0.17	0.27	0.22	0.60	0.10	0.68
21	0.70	0.42	0.62	0.40	0.39	0.36	0.88
22	1.65	2.80	0.91	0.93	0.40	0.35	0.99

Source: Rowland's Analysis, 2021

2.0 Stage 1: Cluster Analysis in Excel

This section essentially deals with the analysis of the data using Microsoft Excel 365. As highlighted previously, it began with normalizing the raw data, by diving each of the concentrations by the mean of the sum total of that particular concentration to obtain the data shown in Figure 2. Thereafter, the distance matrix was calculated in excel using the Euclidean distance Matrix formular given by;

$$\sqrt{\sum_{j=1}^{j=n}\left(x_i - y_i\right)^2}$$

Where;

x_i = values of concentrations at first point of interest

y_i = values of concentrations at second point of interest

j = all the nutrients/cations considered for all pairs of x and y

Series of calculations were carried out. All the distances between pairs of points where nutrients measurements were made was calculated. The result of the calculations shown in Table 3 reveals the various distances between all measurement points.

Table 3: Distance Matrix between Nutrients Measurement Points

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	0																					
2	1.283706	0																				
3	2.095775	2.384696	0																			
4	1.147167	0.617783	2.432317	0																		
5	7.648378	7.768824	5.771670	7.954822	0																	
6	1.974131	2.180403	0.557629	2.222311	6.064130	0																
7	1.115844	0.352376	2.369155	0.584376	7.739846	2.180188	0															
8	6.574588	6.954550	4.843695	6.947522	5.214656	4.836721	6.986676	0														
9	1.995126	0.941897	2.773028	1.112401	7.971658	2.450324	1.072968	7.019659	0													
10	1.134907	1.271078	1.416080	1.437582	6.899462	1.303787	1.326258	5.959350	1.886467	0												
11	2.382295	2.258285	1.297274	2.437279	6.208465	1.405878	2.417305	5.305635	2.681127	1.354739	0											
12	3.484301	3.755171	1.574366	3.799943	5.137889	1.829356	3.817271	3.699394	4.070341	2.691993	1.797237	0										
13	1.503099	0.388674	2.358017	0.854135	7.624641	2.170845	0.668054	6.911367	1.007636	1.219858	2.078151	3.664716	0									
14	2.075668	0.889484	2.921254	1.027708	8.074341	2.698400	1.059151	7.363437	0.705874	1.979666	2.679257	4.205465	0.850318	0								
15	7.514357	7.022393	7.425171	7.272111	10.148020	7.440281	7.314639	9.686431	7.265015	6.760437	6.192716	7.249751	6.700369	6.993394	0							
16	3.458391	2.824752	3.210766	3.206471	7.562751	3.030830	3.138398	6.335182	2.898982	2.572755	2.171867	3.520649	2.600233	2.993812	5.151252	0						
17	1.636155	0.870684	2.795072	0.976914	8.040488	2.576670	0.859235	7.282899	0.863722	1.744339	2.766208	4.172693	0.992888	0.964748	7.307180	3.188465	0					
18	3.505391	3.194472	3.663058	3.668681	7.810113	3.443631	3.305783	6.762546	3.251539	3.129862	3.503335	4.418636	3.241962	3.661100	7.522424	2.809605	3.383356	0				
19	1.863183	0.794406	2.801908	1.040159	7.978887	2.576482	0.968718	7.250658	0.720301	1.709933	2.555769	4.092194	0.711508	0.625050	6.849986	2.811386	0.561070	3.386226	0			
20	1.855798	0.745867	2.844364	0.918000	8.048416	2.609111	0.878129	7.307344	0.594679	1.809664	2.668741	4.172473	0.749592	0.422120	7.045830	2.991116	0.573386	3.494798	0.290894	0		
21	1.423626	0.465886	2.276779	0.646779	7.645852	2.025748	0.556528	6.764754	0.659183	1.334385	2.256151	3.649534	0.617890	0.814937	7.203300	2.880806	0.733846	3.248824	0.716667	0.641290	0	
22	2.934146	2.808432	2.540252	2.961719	7.060549	2.021782	2.891291	5.017154	2.623062	2.458088	2.745788	3.205834	2.827243	3.175828	7.646172	2.891035	3.090459	2.837612	3.032066	3.062339	2.637722	0

Source: Rowland's Analysis, 2021

From Table 3, it can be observed that, the minimum distance between all pairs of measurement points is 0.290894 (as highlighted in yellow colour), which occurred at point 19 and 20 (that is, column 19 and row 20). As a result, points 19 and 20 were combined to form a new cluster called 19* to replace both 19 and 20 on the rows and columns (vertical and horizontal axis). Medium linkage method is used in this excel version, and this was achieved by calculating the average of the sum of distances of points between 19 and 20 to obtain the new distance for 19*. The reduced similarity matrix, that includes 19* is therefore depicted in Table 4.

The similarity matrix was further reduced by identifying the next point with the lowest distance, which eventually equals 0.352376, and falls between points 2 and 7 (that is, column 2 and row 7), as highlighted in yellow colour in Figure 4. This points were merged into a new cluster called 2* to represent the average distances between points 2 and 7 in the matrix. The new reduced matrix that contains both 2* and 19* is therefore depicted in Table 5. The Distance matrix was further reduced to obtain the third level reduced matrix by following the same procedure discussed above.

Table 4: Reduced Similarity Matrix 1: Where 19* = (19 & 20)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19*=(19 & 20	21	22
1	0						1.115844														
2	1.283706	0					0.352376														
3	2.095775	2.384696	0				2.369155														
4	1.147167	0.617783	2.432317	0			0.584376														
5	7.648378	7.768824	5.771670	7.954822	0		7.739846														
6	1.974131	2.180403	0.557629	2.222311	6.064130	0	2.180188														
7	1.115844	0.352376	2.369155	0.584376	7.739846	2.1801882	0														
8	6.574588	6.954550	4.843695	6.947522	5.214656	4.8367207	6.986676	0													
9	1.995126	0.941897	2.773028	1.112401	7.971658	2.4503238	1.072968	7.019659	0												
10	1.134907	1.271078	1.416080	1.437582	6.899462	1.3037873	1.326258	5.959350	1.886467	0											
11	2.382295	2.258285	1.297274	2.437279	6.208465	1.4058781	2.417305	5.305635	2.681127	1.354739	0										
12	3.484301	3.755171	1.574366	3.799943	5.137889	1.8293556	3.817271	3.699394	4.070341	2.691993	1.797237	0									
13	1.503099	0.388674	2.358017	0.854135	7.624641	2.1708451	0.668054	6.911367	1.007636	1.219858	2.078151	3.664716	0								
14	2.075668	0.889484	2.921254	1.027708	8.074341	2.6983998	1.059151	7.363437	0.705874	1.979666	2.679257	4.205465	0.638706	0							
15	7.514357	7.022393	7.425171	7.272111	10.148020	7.4402806	7.314639	9.686431	7.265015	6.760437	6.192716	7.249751	6.718053	6.993394	0						
16	3.458391	2.824752	3.210766	3.206471	7.562751	3.0308300	3.138398	6.335182	2.898982	2.572755	2.171867	3.520649	2.615189	2.993812	5.151252	0					
17	1.636155	0.870684	2.795072	0.976914	8.040488	2.5766700	0.859235	7.282899	0.863722	1.744339	2.766208	4.172693	1.116581	0.964748	7.307180	3.188465	0				
18	3.505391	3.194472	3.663058	3.668681	7.810113	3.4436311	3.305783	6.762546	3.251539	3.129862	3.503335	4.418636	3.265439	3.661100	7.522424	2.809605	3.383356	0			
19*=(19 & 20)	1.859490	0.770137	2.823136	0.979079	8.013651	2.5927962	0.923423	7.279001	0.657490	1.759799	2.612255	4.132333	0.730550	0.523585	6.947908	2.901251	0.567228	3.440512	0		
21	1.423626	0.465886	2.276779	0.646779	7.645852	2.0257481	0.556528	6.764754	0.659183	1.334385	2.256151	3.649534	0.836540	0.814937	7.203300	2.880806	0.733846	3.248824	0.678978	0	
22	2.934146	2.808432	2.540252	2.961719	7.060549	2.0217825	2.891291	5.017154	2.623062	2.458088	2.745788	3.205834	2.939124	3.175828	7.646172	2.891035	3.090459	2.837612	3.047202	2.637722	0

Source: Rowland's Analysis, 2021

Table 5: Reduced Similarity Matrix 1: Where 2* = (2 & 7)

	1	2*=(2&7)	3	4	5	6	8	9	10	11	12	13	14	15	16	17	18	19*=(19 & 2	21	22
1	0																			
2*=(2&7)	1.199775	0																		
3	2.095775	2.376925	0																	
4	1.147167	0.601079	2.432317	0																
5	7.648378	7.754335	5.771670	7.954822	0															
6	1.974131	2.180296	0.557629	2.222311	6.064130	0														
8	6.574588	0.176188	4.843695	6.947522	5.214656	4.836721	0													
9	1.995126	6.970613	2.773028	1.112401	7.971658	2.450324	7.019659	0												
10	1.134907	1.007433	1.416080	1.437582	6.899462	1.303787	5.959350	1.886467	0											
11	2.382295	1.298668	1.297274	2.437279	6.208465	1.405878	5.305635	2.681127	1.35474	0										
12	3.484301	2.337795	1.574366	3.799943	5.137889	1.829356	3.699394	4.070341	2.69199	1.797237	0									
13	1.503099	3.786221	2.358017	0.854135	7.624641	2.170845	6.911367	1.007636	1.21986	2.078151	3.664716	0								
14	2.075668	0.528364	2.921254	1.027708	8.074341	2.698400	7.363437	0.705874	1.97967	2.679257	4.205465	0.638706	0							
15	7.514357	0.974317	7.425171	7.272111	10.148020	7.440281	9.686431	7.265015	6.76044	6.192716	7.249751	6.718053	6.993394	0						
16	3.458391	7.168516	3.210766	3.206471	7.562751	3.030830	6.335182	2.898982	2.57275	2.171867	3.520649	2.615189	2.993812	5.151252	0					
17	1.636155	2.981575	2.795072	0.976914	8.040488	2.576670	7.282899	0.863722	1.74434	2.766208	4.172693	1.116581	0.964748	7.307180	3.188465	0				
18	3.505391	0.864960	3.663058	3.668681	7.810113	3.443631	6.762546	3.251539	3.12986	3.503335	4.418636	3.265439	3.661100	7.522424	2.809605	3.383356	0			
19*=(19 & 20)	1.859490	3.250127	2.823136	0.979079	8.013651	2.592796	7.279001	0.657490	1.75980	2.612255	4.132333	0.696988	0.523585	6.947908	2.901251	0.567228	3.440512	0		
21	1.423626	0.846780	2.276779	0.646779	7.645852	2.025748	6.764754	0.659183	1.33439	2.256151	3.649534	0.836540	0.814937	7.203300	2.880806	0.733846	3.248824	0.716667	0	
22	2.934146	0.511207	2.540252	2.961719	7.060549	2.021782	5.017154	2.623062	2.45809	2.745788	3.205834	2.939124	3.175828	7.646172	2.891035	3.090459	2.837612	3.032066	2.637722	0

Source: Rowland's Analysis, 2021

Stage 2: Cluster Analyses Using R Studio

The second aspect of this analysis employs R-Studio version 4.2.1 to analyse the data set with the aid of the blocks of R codes provided by the course lecturer.

3.1 Preparation of the Data for Analysis in R

The normalised data in excel (that contains concentrations divided by their averages) were converted into .csv (comma delimited) file, and consequently loaded into RStudio after setting the environment through the use of the blocks of codes below;

```
```{r}
setwd("C:/Environmental_Monitoring_Project_2") #Setting the Environment
data=read.table("Cluster_New.csv", header=TRUE, row.names=1, dec=".", sep=",") #Reading the Table
library(factoextra) #Loading factoextra library that'll be used later
```
```

Source: Adapted from Environmental Monitoring Project Material by Loga Małgorzata, 2021

3.2 Calculating Distances and Plotting of Dendrograms using all Available Methods

After successfully loading the data, the distance matrix, and clusters were calculated using all the available distance methods and clustering techniques provided in the sample code. For Euclidean method, *the distance calculated with RStudio was entirely the same with the distance calculated manually using excel*. A sample of the code that was used for calculating the Euclidean distance, with ward.D clusters is provided below;

```
```{r}
dismat<-dist(data, method = "euclidean") # Calculating the distance, using Euclidean method
cluster1<-hclust(dismat, method = "ward.D") #Using ward.D method of Clustering
plot(cluster1, cex = 0.5)
dev.copy(png,"Euclidean_Plots/Euclidean_with_Wardd.png", units='in', width=8.5,height=6,res=1200)
#Saving the Image
```
```

Source: Adapted from Environmental Monitoring Project Material by Loga Małgorzata, 2021

This was repeated for the all methods including Maximum, Manhattan, Canberra, Binary and Minkowski. In addition, for each of the methods, the following clustering methods were used; Ward.D, Ward.D2, Single, Complete, Average, Mcquitty, Median and Centroid. This was done in order to have variety of options so as to make appropriate selection on clusters based on the groupings that appears most.

The various plots obtained from the different combinations of distance methods and clustering techniques applied in this analyses are therefore provided in the next section 3.4, while the full versions of the codes (which includes all distance and grouping methods used) are also included in the Appendix.

3.3 Overview of the Clustering Methods Applied in this Analyses

A) The Ward.D Clustering Method: In statistics, Ward's method is a method of clustering applied in hierarchical cluster analysis. The procedure involves agglomerative hierarchical clustering procedure, where the criterion for choosing the pair of clusters to merge at each step is based on the optimal value of an objective function. The method is basically seeks to create groups in such a way that variance between points of measurement is minimized within clusters (University of Alberta, Year (NA)).

B) The Ward.D2 Clustering Method: The Ward.D2 is an improvement on the ward.D, the algorithm used here is similar to that of Ward.D, the only difference is that, here, dissimilarities are squared before clustering is done.

C) Single Linkage Clustering Method: This method approaches the distance between two clusters as the minimum distance between their members. It is referred to as “single link” due to the fact that it specifies that clusters are close if they have even a single pair of close points, which is called a single “link”. This can handle quite complicated cluster shapes. The approach only considers separation, at the detriments of compactness or balance (Carnegie Mellon University, 2009).

D) Complete Linkage Clustering Method: In this method, the distance between clusters is the maximum distance between their members. This method does not always work well for all data, For some, it performs excellently, while for others, it fails outrightly (Carnegie Mellon University, 2009).

E) Average linkage Clustering Method: In this method, the distance between each pair of observations in each cluster are summed up and divided by the number of pairs to obtain an average inter-cluster distance (Clemens, 2019).

F) Centroid Linkage Clustering Method: This involves the distance between centroids of two clusters. This method is based on the principle that as centroids move with new observations, there is likelihood that the smaller clusters will be more similar to the new larger cluster than to their individual clusters, thereby causing an inversion in the dendrogram.

G) Mcquitty Clustering Method: This method involves calculating the distance of a new cluster to any other cluster by taking the average of the distances between the clusters. It is particularly useful when two clusters are to be joined together (University of Alberta, (Year (NA)).

H) Median Clustering Method: Unlike the average linkage, the median clustering attempt to minimize the distance between points in a cluster and a point designated as the center of that cluster.

3.4 Dendrograms Generated Using Euclidean Distance Method

In this section, the Euclidean distance method was implemented, with the various linkage methods discussed above in order to determine the clusters in the dataset, which will consequently help to determine the areas (represented by the measurement points) that indicate the common toll of penguins in the study domain.

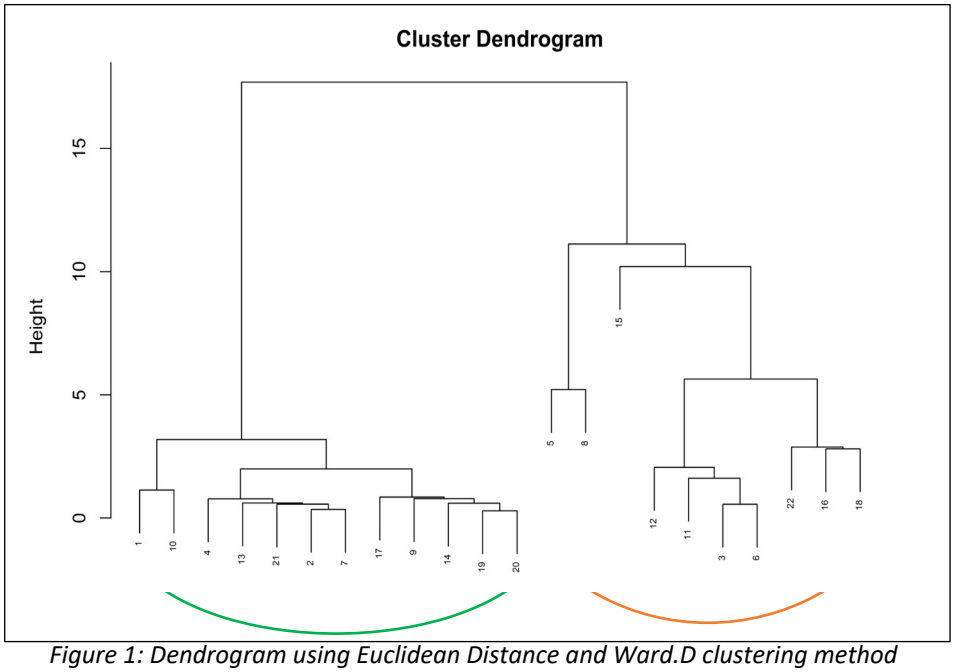


Figure 1: Dendrogram using Euclidean Distance and Ward.D clustering method

Clusters depends on the level of similarity of a data, expressed by the distances between the individual points in that data. There are several conflicting ideas and approaches regarding determination of clusters from dendrograms - it is to a great extent subjective, and based on intuition. However, the various approaches agrees that a distance threshold should be determined on the dendrogram at which, any distance above will not be considered. As Krish (2019) noted, the number of clusters in a particular dendrogram can be determined by using the rule of the vertical distances between points – that is, by considering the longest vertical line (with the highest distance), such that none of the other horizontal line can pass through it (as shown with the dotted red line in Figure 1). James (2015), also shares similar views, but greatly believes that two clusters is never an ideal or perhaps a meaningful solution, when an array of data is involved. However, following the Krish (2019) method, in Figure 1, there appears to be *two main clusters* in the groups. The first obvious cluster consist of points 1, 10, 4, 13, 21, 2, 7, 17, 9 14, 19 and 20, while the second consists of points 5, 8, 15, 12, 11, 15, 3, 6, 22, 16, and 18. Although, if critically examined, it can be argued that points 5 and 8, have the tendency to form a separate cluster, and that point 15, stands separately alone. Hence, the Krish (2019) method might not be applicable for all cases. However, for simplicity it might be save to consider these two points (5 and 8) as part of the larger second group. Figure 1, also shows that there are multiple levels of clusters in the data; the first main cluster contain three sub-clusters, while the second main cluster contain arguably four sub-clusters. Hence, the reason for the name hierarchical agglomeration.

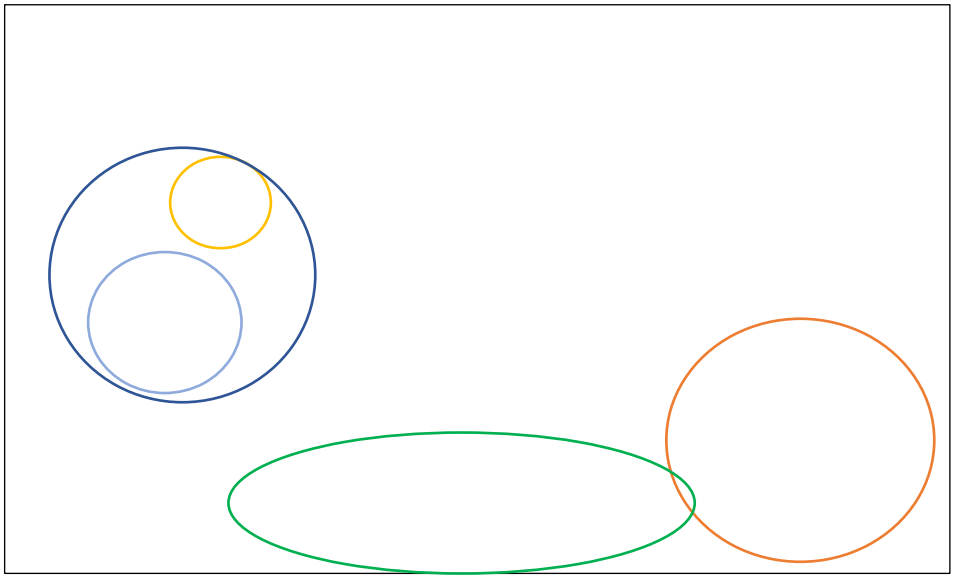


Figure 2: Dendrogram using Euclidean Distance and Ward.D2 clustering method

The Kish (2019) approach does not explicitly apply to the dendrogram result shown in Figure 2, however, it seems obvious that there are about four clusters in the dendrogram (although 3 are obvious using Kish (2019) rule). The first cluster consist of points 5 and 8, the two points having about 5.5 vertical distances apart. Close to this, two points is point 15. The second main cluster includes points 1, 10, 4, 13, 21, 2, 7, 17, 9, 14,19, and 20, while the fourth cluster includes points 12, 11, 3, 6, 22, 16, and 18. The interesting thing in Figure 1, and Figure 2 is that, points 5 and 8, both somewhat forms a separate cluster in the two algorithms, and majority of the elements that are within a cluster in Figure 1, are also somewhat within a separate cluster in Figure 2.

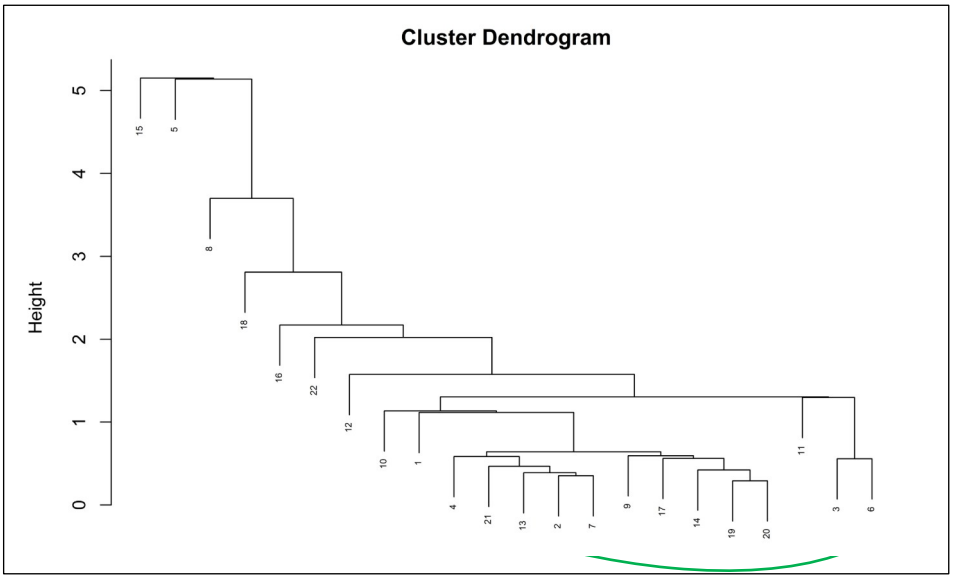


Figure 3: Dendrogram using Euclidean Distance and Single clustering method

The algorithm here exhibits a somewhat strange result, however, it appears that that are two clusters in Figure 3. Points 15 and 5 forms a cluster, while the other remaining points appears to belong to the same cluster, with sub-clusters within. The algorithm here, exhibits a

complex hierarchical agglomeration, with almost all of the points emanating from a particular cluster.

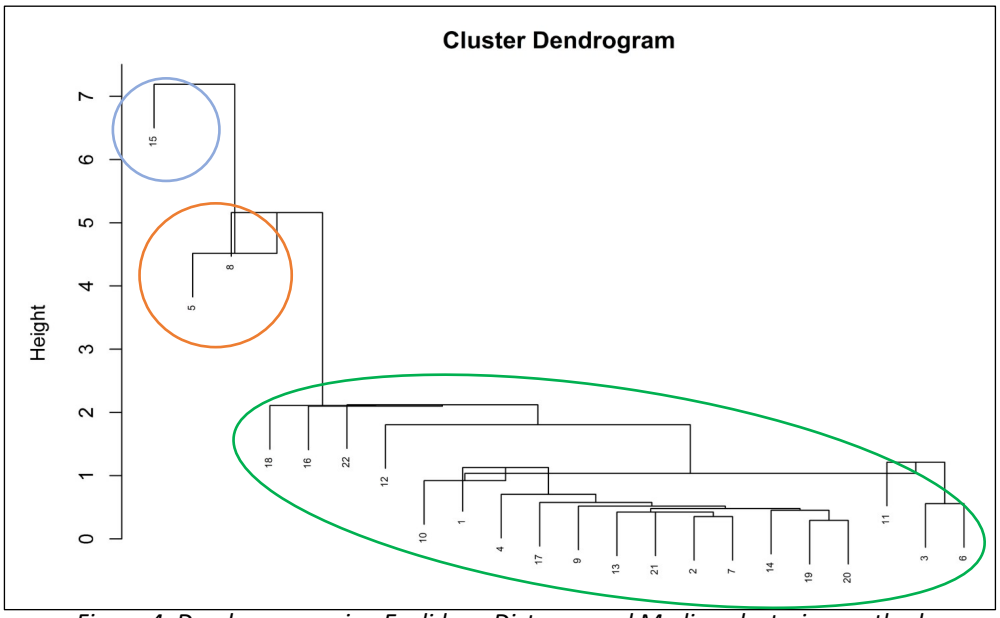


Figure 4: Dendrogram using Euclidean Distance and Median clustering method

The result obtained using the Median clustering exhibits similarities with the results obtained for Single clustering. From Figure 4, it can be seen that point 15 stood alone, while point 5 and 8 forms a cluster. Meanwhile, the remaining points are somewhat within the same cluster. So far, it seems pretty obvious that point 5 and 8 and 15, has some characteristics which sets them apart from other locations. It might be too early to decide anyway, therefore, further analysis of the remaining dendrogram will greatly help to validate this claim.

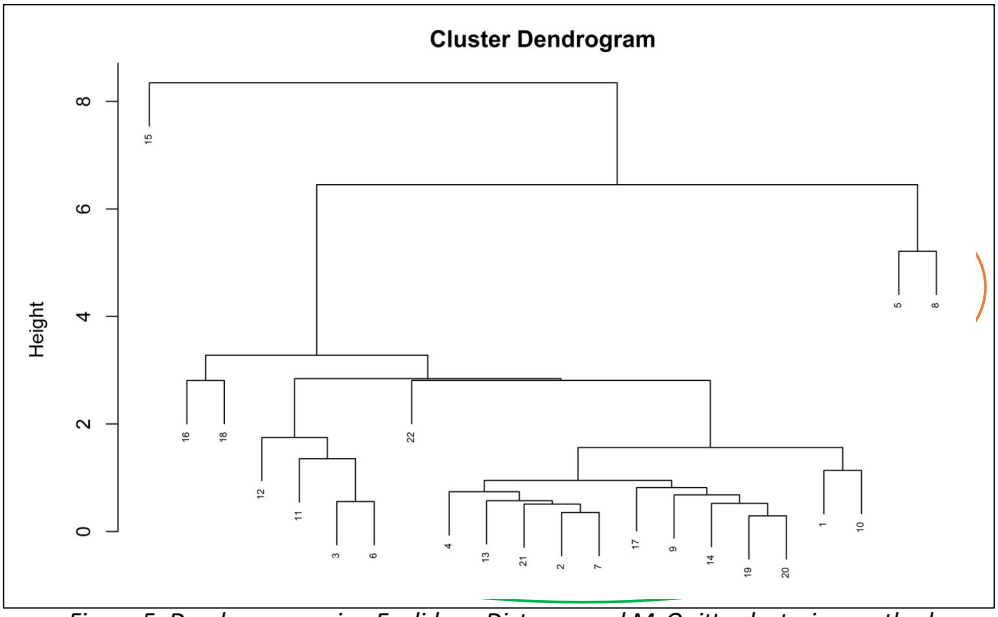


Figure 5: Dendrogram using Euclidean Distance and McQuitty clustering method

Again, using the McQuitty method, as displayed in Figure 5, there also appears to be three main clusters; out of these three, point 15 stands alone, while like some previous plots, points 5 and 8 forms a separate cluster. Meanwhile, the remaining points creates a cluster, to form

the cluster with the highest number of points. Within this cluster, there are also separate smaller groups of about five clusters.

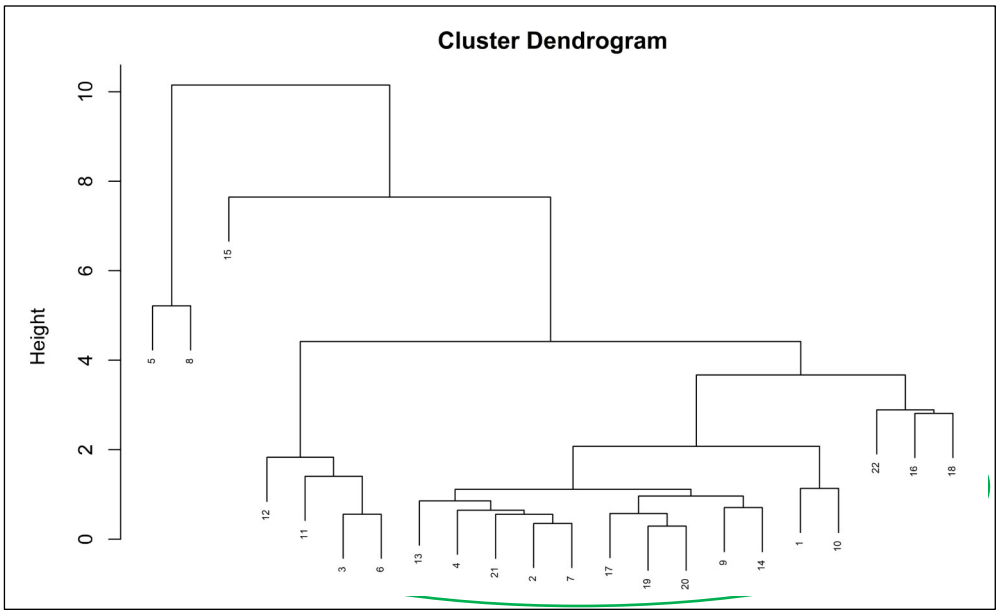


Figure 6: Dendrogram using Euclidean Distance and Complete clustering method

Likewise, for complete clustering method, there appears to be about three main clusters. Again, 5 and 8, forms a cluster, 15 stands alone, and the remaining points form a group of larger cluster, with sub clusters. The result obtained here essentially agrees with that obtained using the McQuitty method.

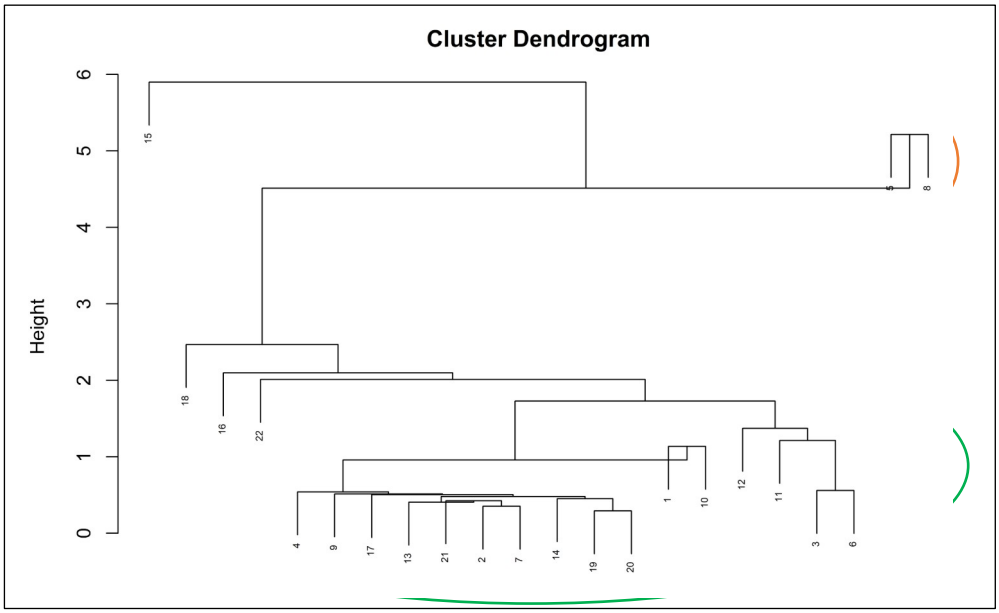


Figure 7: Dendrogram using Euclidean Distance and Centroid clustering method

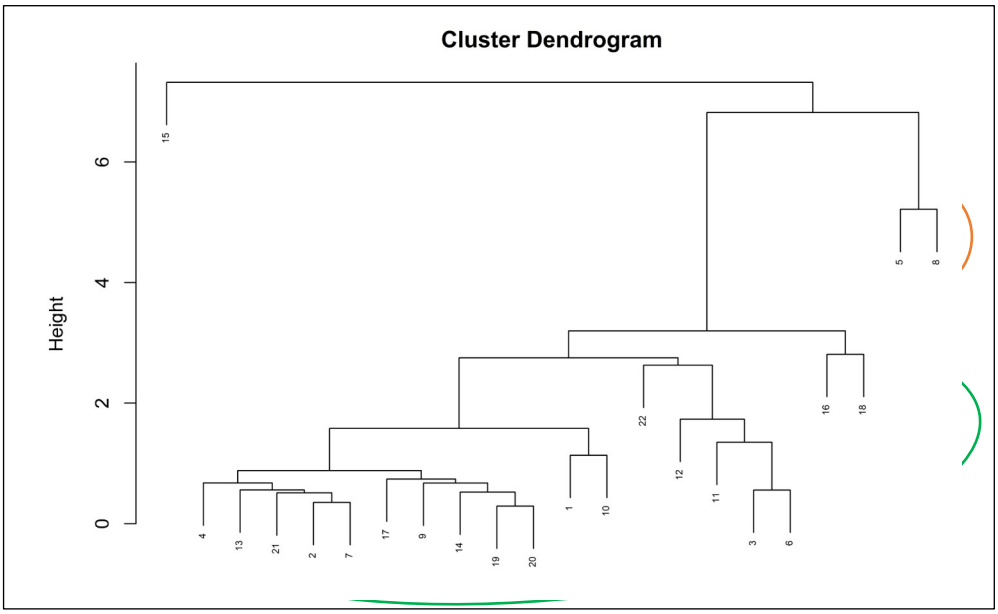


Figure 8: Dendrogram using Euclidean Distance and Average clustering method

The last methods of clustering used within the Euclidean distance methods are Average and Complete methods. The dendrograms generated using the two methods also exhibits similar characteristics when compared to the other dendrogram plots obtained using other methods. From Figure 7 and 8, it is obvious that there are three main clusters in the two plots. From the two figures, point 15 stands alone, while points 5 and 8 forms a cluster, whereas the remaining points converge together to form a larger cluster, containing sub clusters that are lower in hierarchy. From the various algorithms used, it is pretty obvious that with the various clustering adopted for Euclidean distance, the most repeated number of clusters in three (3). However, attempts can still be made to further examine the possible alternatives for grouping the data using other distance methods.

3.5 Dendrograms Generated Using Manhattan Distance Method

The Manhattan Distance is the sum of absolute differences between points across all the dimensions, and as such, it was served as an invaluable tool in determined the number of probable clusters in this analysis. It was also used with other clustering methods to determine the number of available clusters in the dataset.

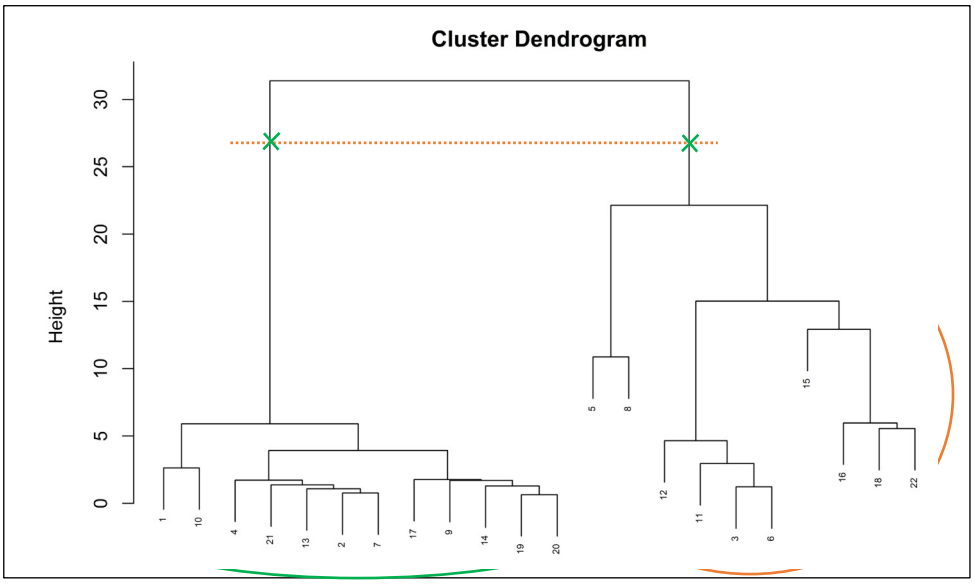


Figure 9: Dendrogram using Manhattan Distance and Ward.D clustering method

There are two obvious clusters in Figure 9. The first cluster includes measurement points 1, 10, 4, 21, 13,2, 7, 17, 9, 14 19 and 20. While the second group includes measurement points 5, 8, 12, 11, 3, 6, 15, 16, 18 and 22. However, if one takes a critical look at the dendrogram, it could be argued that measurement point 5 and 8 appears to form a distinct cluster – thereby making it the third probable cluster obtained with this algorithm. The result obtained here somewhat agrees with some of the results earlier analysed.

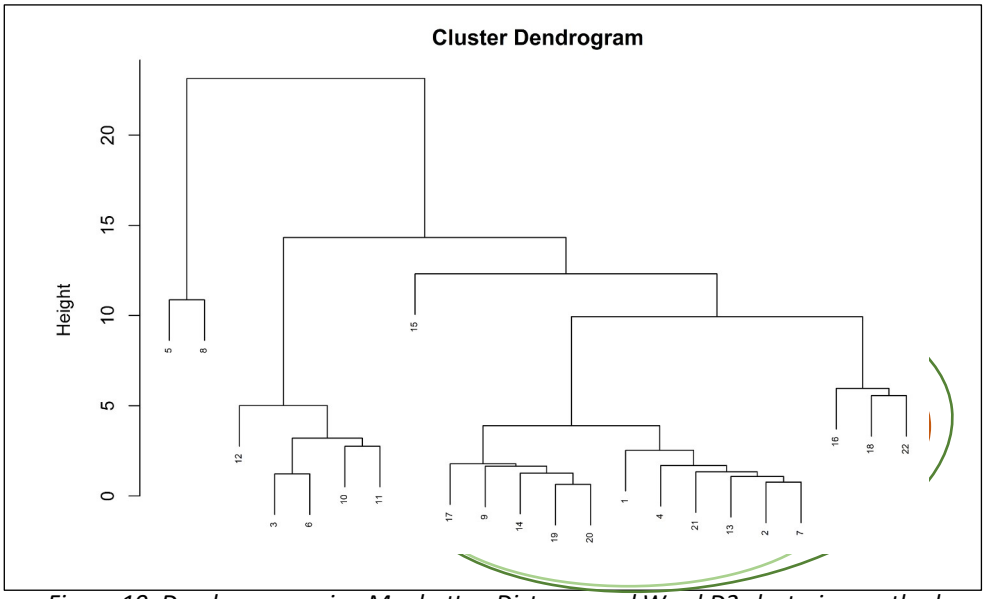


Figure 10: Dendrogram using Manhattan Distance and Ward.D2 clustering method

From Figure 10, it is obvious that there are at least four clusters in the dendrogram. Perhaps, the Interesting thing about this dendrogram is that, despite that the number of clusters here is somewhat different from the others earlier analysed, nevertheless, there still exist some similarities with the previously analysed dendrograms. Notably, like other dendrograms, measurements points 5 and 8 forms a cluster, while measurement 15 stands alone. 12, 3, 6, 10, 11, forms a cluster, while the remaining measurement points formed another cluster, which can further be divided into two groups of clusters.

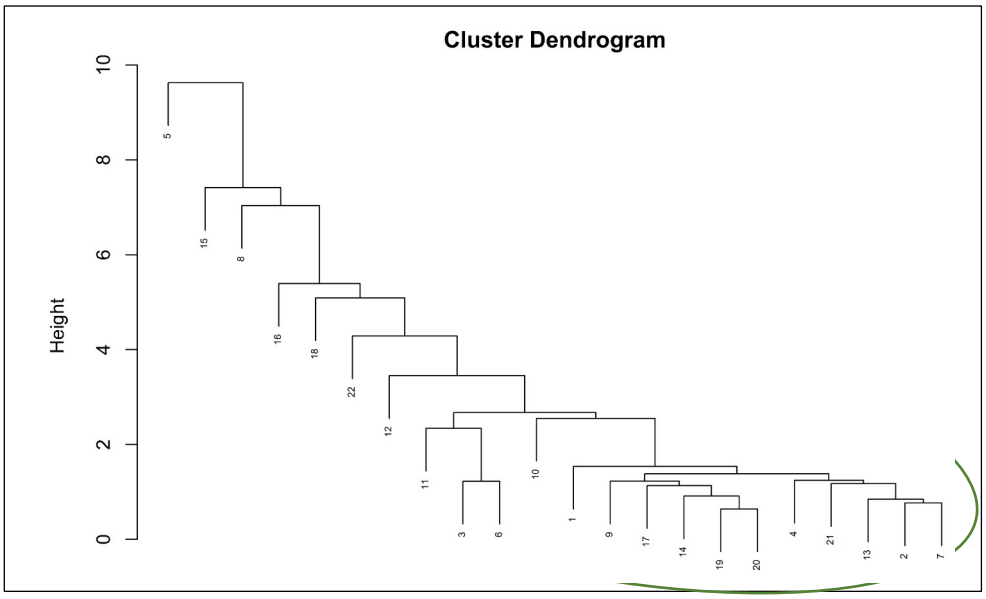


Figure 11: Dendrogram using Manhattan Distance and Single-Linkage clustering method

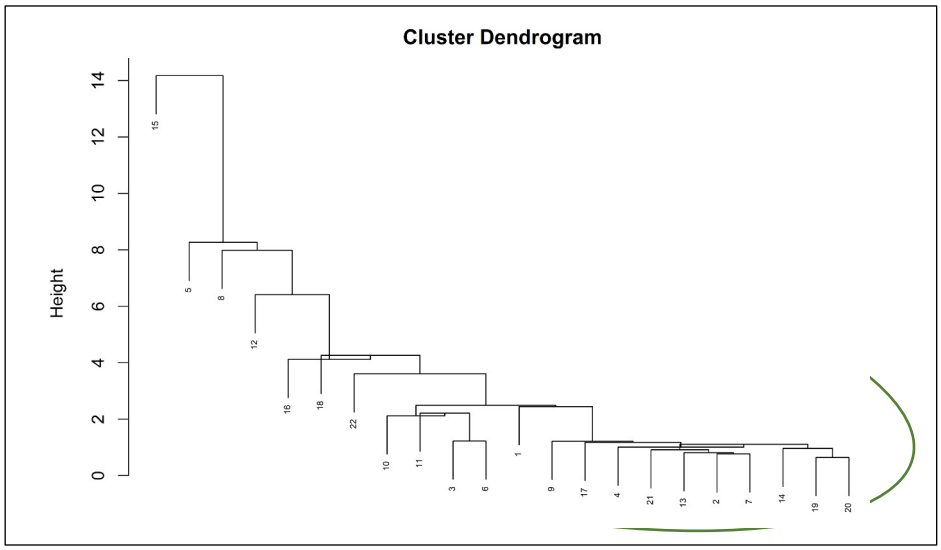


Figure 12: Dendrogram using Manhattan Distance and Median clustering method

The groupings provided by the algorithm depicted in Figure 11 and Figure 12, appears to be inelegant. At first observation, it appears that the measurement points appears to all belong to a single main cluster with some sub groups. The result here is ambiguous, and somewhat related to the result obtained for the Single-linkage and Median methods using with Euclidean distance algorithm. However, a closer observation at the plots reveals that, there are three clusters in Figure 16. Point 5 stands alone, points 15 and 8 somewhat forms a cluster, while

the remaining points forms a larger cluster. Meanwhile, for Figure 16, there appears to be two clusters. Point 15 stands alone, while the remaining point somewhat forms a cluster with sub-groups.

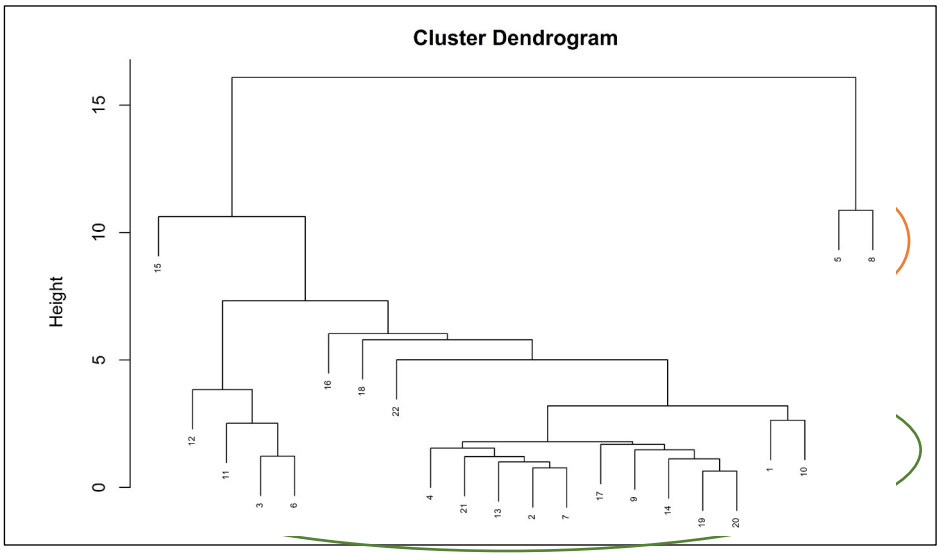


Figure 13: Dendrogram using Manhattan Distance and McQuitty clustering method

The dendrogram depicted in Figure 13 made use of the Manhattan Distance with McQuitty linkage method. From the dendrogram, there are arguably three clusters in the dataset. The interesting thing is that Like the majority of the results of other algorithms, measurement points 5 and 8 forms a cluster, while the remaining measurements points appears to be greatly linked to one another, though, with sub groups; points 12, 11, 3, and 6 belongs to a sub-cluster, while, the remaining measurement points belongs to another cluster, except point 15 which somewhat stands alone, and hierarchically higher than the other clusters, but somewhat at the same level with measurement points 5 and 8.

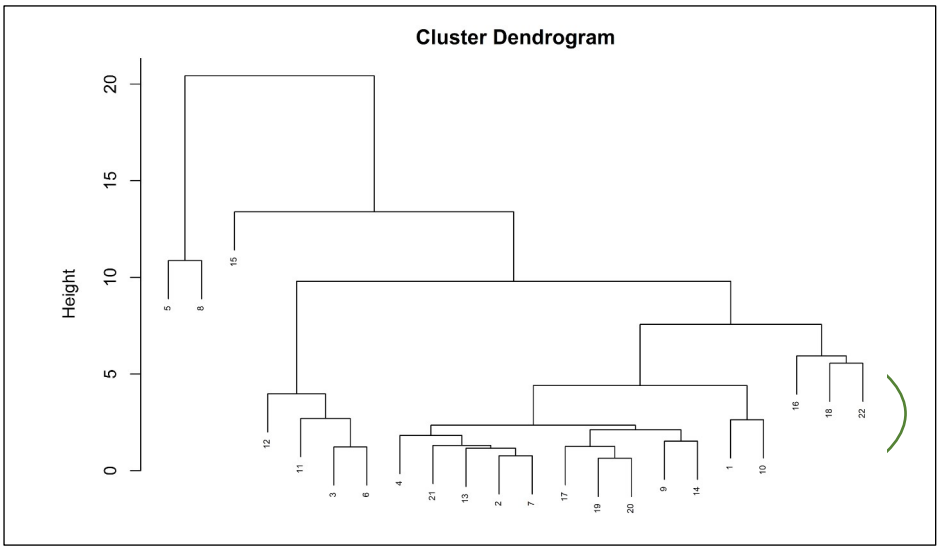


Figure 14: Dendrogram using Manhattan Distance and Complete-Linkage clustering method

Similar result is also seen in Figure 14, which is built on the premise of Manhattan Distance algorithm with Complete-Linkage clustering method. Still, points 5 and 8 formed a separate

cluster, while the remaining points appears to belong to a larger cluster, with point 15 standing alone and higher than other points in hierarchy.

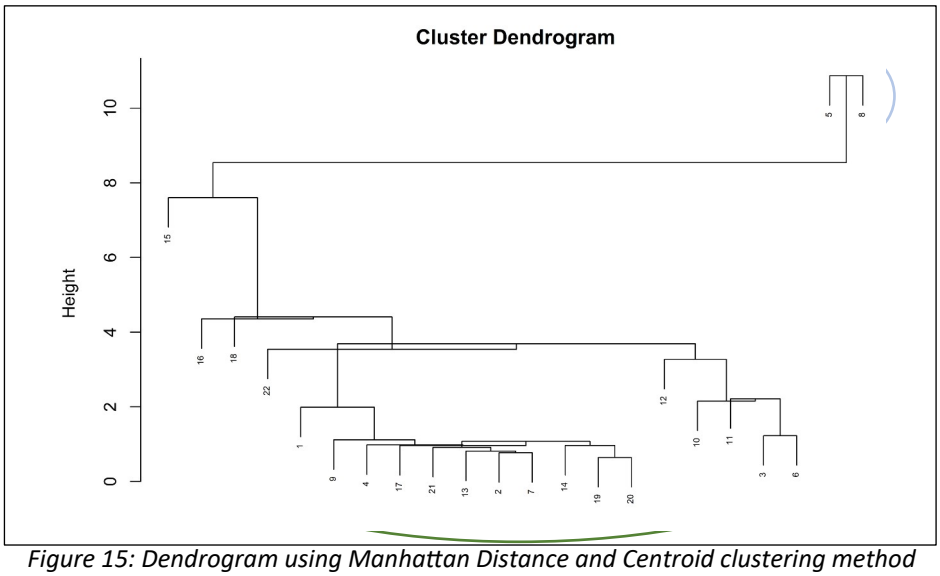


Figure 15: Dendrogram using Manhattan Distance and Centroid clustering method

There are about three clusters in Figure 15, the first group includes measurement points 5 and 8, the second includes point 15 standing almost alone, while the third group includes the agglomeration of the remaining measurement points, The result obtained here, is also somewhat similar to the results obtained for the previous algorithms.

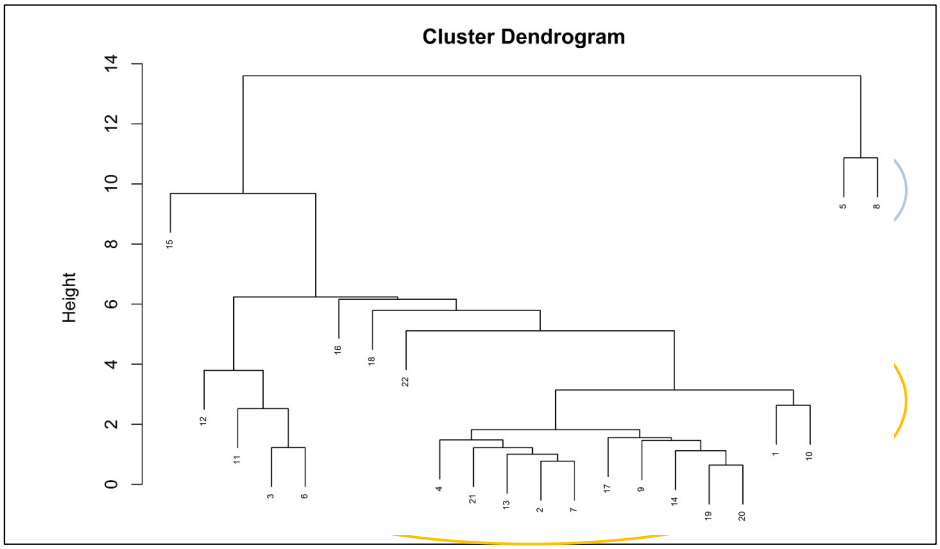


Figure 16: Dendrogram using Manhattan Distance and Average-Linkage clustering method

Finally for Manhattan Distance method, the average-linkage clustering method was used, and the result depicted in Figure 16 revealed that, there are also about three clusters with the same contents with the one obtained in Figure 15. The obvious cluster however, is the sperate cluster formed by measurement points 5 and 8, and another one formed by measurement points 15. Other points are agglomerated together, though could still be further subdivided into separate clusters.

3.6 Dendrograms Generated Using Maximum Distance Method

In this section, the maximum distance method was implemented, with all the available linkage clustering methods. The results of the algorithms, are therefore provided, and discussed using the following dendrograms.

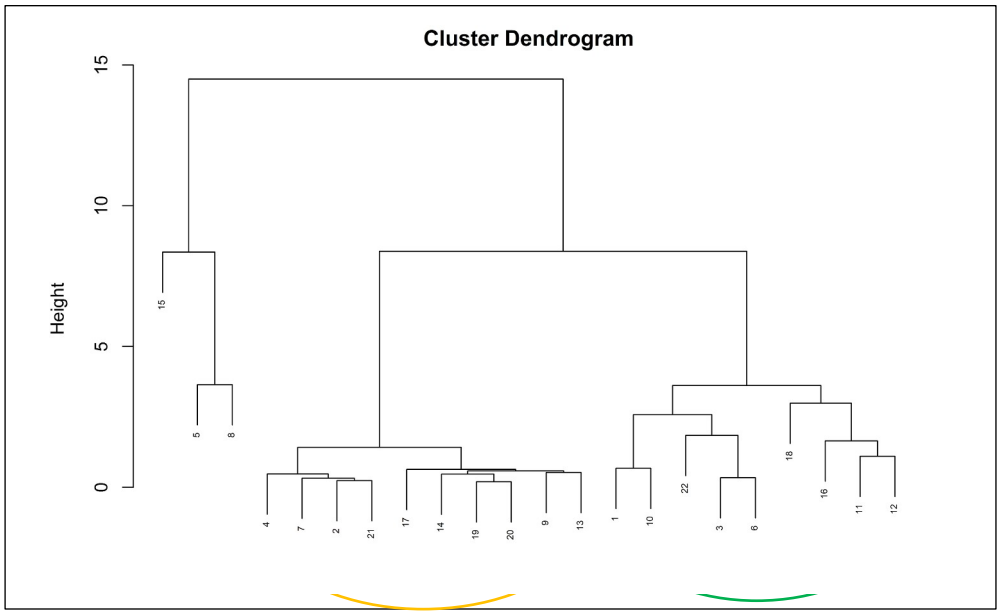


Figure 17: Dendrogram using Maximum Distance and Ward.D clustering method

There are three distinct clusters in the dendrogram depicted in Figure 17. The first cluster involves measurement points 15, 5 and 8. The second cluster includes measurement points 4, 7, 2, 21, 17, 14, 19, 20, 9, and 13, while the last cluster includes points 1, 10, 22, 3, 6, 18, 16, 11, and 12. The interesting thing here is that, the main three groups of clusters are easily identified, and the result obtained with this algorithm appears to be reasonable, as it also agrees with some of the results of the dendrograms earlier analysed in this report.

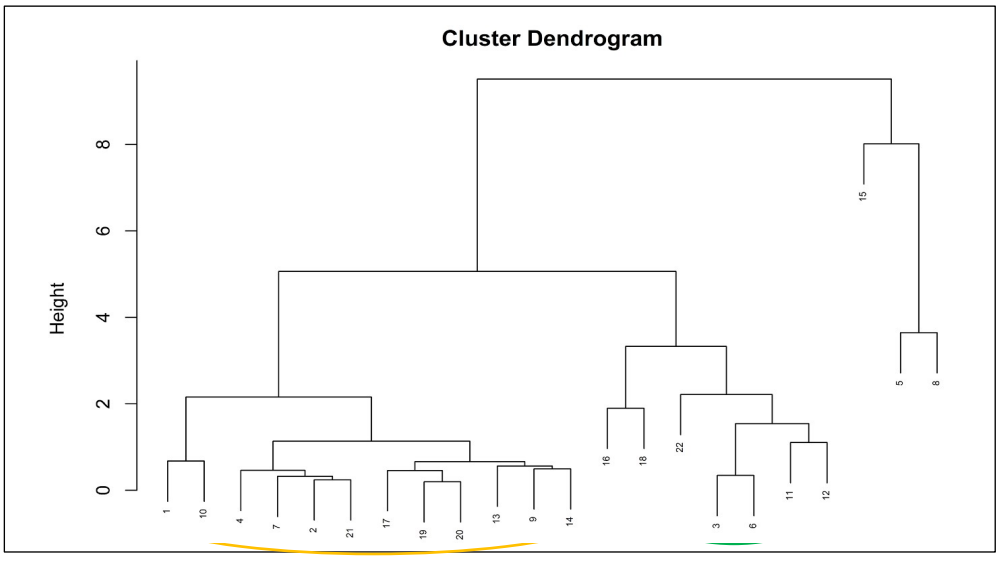


Figure 18: Dendrogram using Maximum Distance and Ward.D2 clustering method

Like in Figure 17, the algorithm implemented in Figure 18 produced an obvious grouping of the dataset. Three groups of clusters are evident here too. Interestingly, points 15, 5 and 8 forms a cluster; though point 15 standing significantly higher than the other two points. while the remaining points forms two clusters. The measurement points included in these two

cluster are not entirely the same with those included in Figure 17. Here, points 1, 10, 4, 7, 2, 1, 17, 19, 20, 13 9 and 14 form belongs to a group, while points 16, 18, 22, 3, 6, 11, and 12 belongs to another group. Though with slight difference, the two algorithms still produces similar composition of the clusters in terms of elements included.

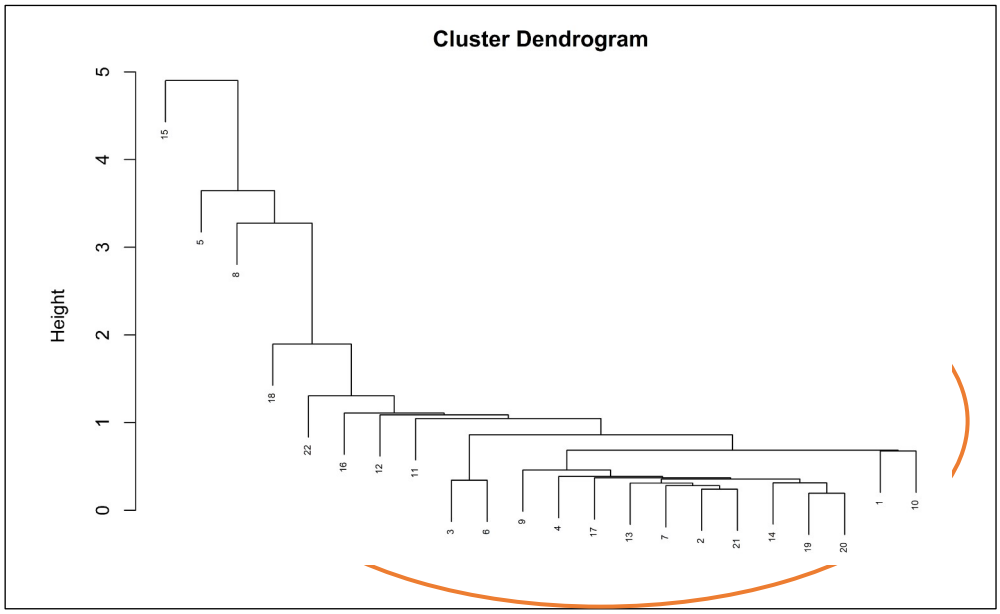


Figure 19: Dendrogram using Maximum Distance and Single-Linkage clustering method

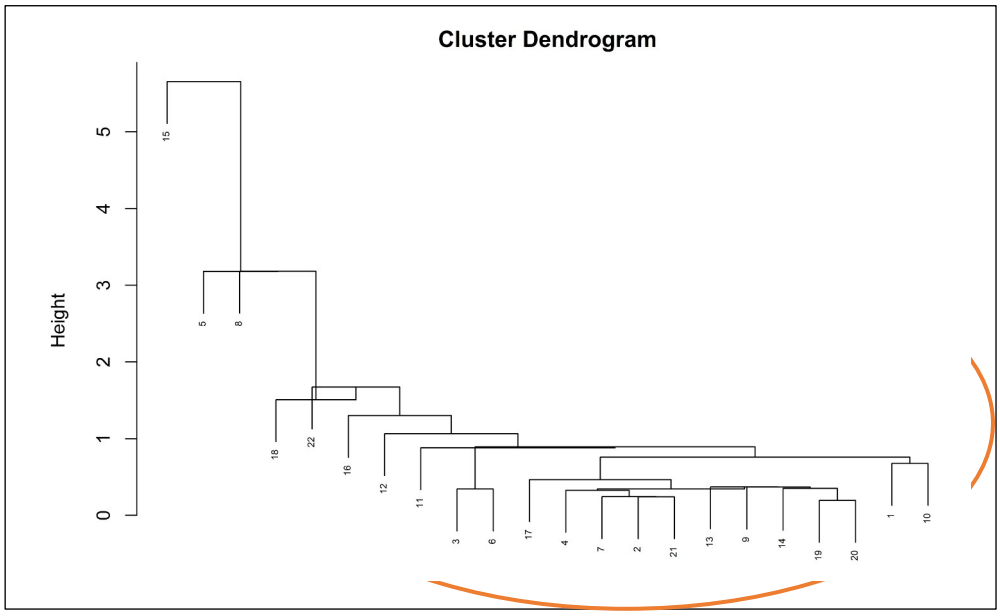


Figure 20:: Dendrogram using Maximum Distance and Median clustering method

The results obtained here using Single Linkage and Median clustering method with Maximum distance is similar to the results previously obtained for the two linkage methods when combined with other distance algorithms. However, from Figure 19 and 20, it appears that there are three clusters, measurement points 5 and 8, formed a group, while point 15 stands

alone, though, still closer to points 5 and 8 (especially in Figure 20). Other measurement points converge to form a big cluster, with many sub-clusters.

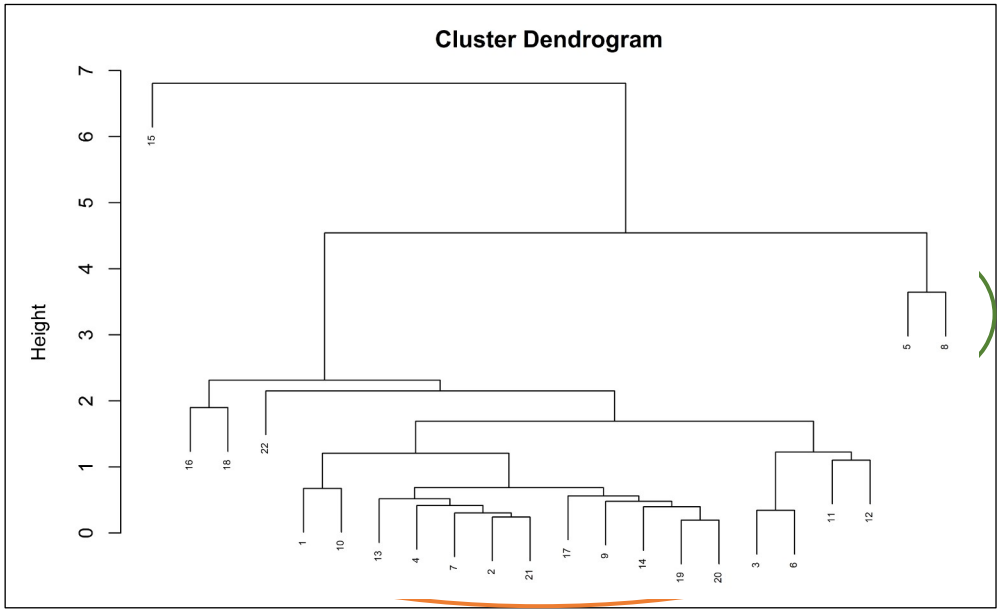


Figure 21: Dendrogram using Maximum Distance and McQuitty clustering method

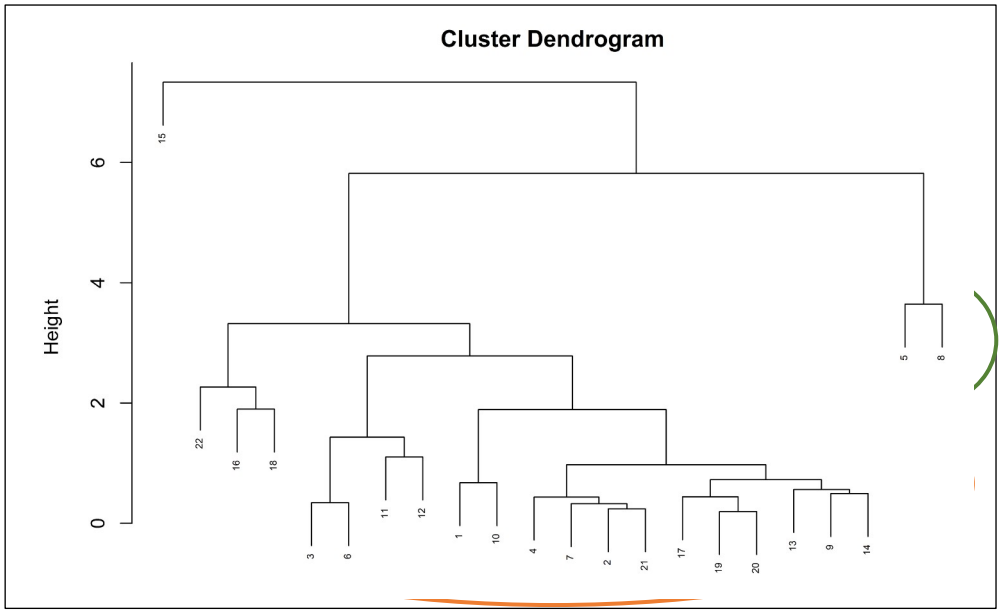
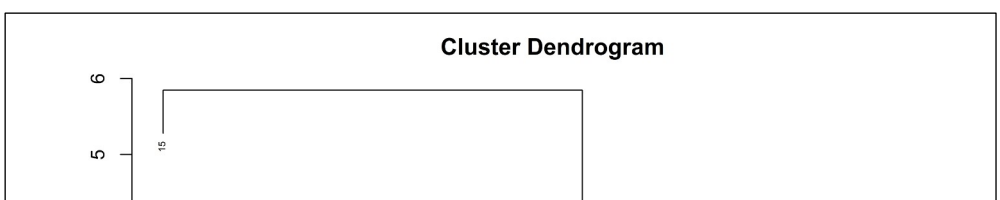


Figure 22: Dendrogram using Maximum Distance and Complete-Linkage clustering method

The results obtained using the McQuitty, and Complete-Linkage with Maximum distance algorithm is particularly interesting! The same results of groupings were obtained with the two linkage methods. As shown in Figure 21 and 22, in both methods, point 15 stands alone to form a separate cluster. As expected, measurement points 5 and 8 stands together to form another cluster, while the remaining measurement points converge to form a large cluster with about six to seven micro-clusters.



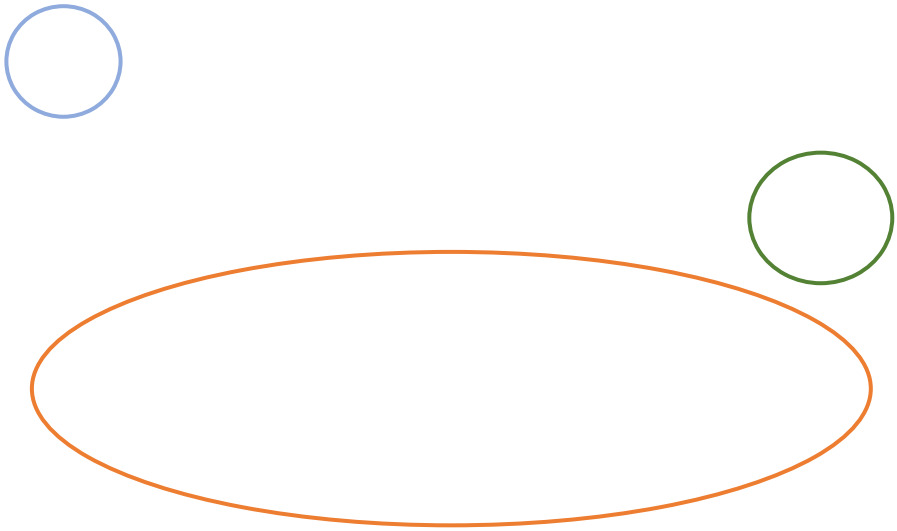


Figure 23: Dendrogram using Maximum Distance and Centroid clustering method

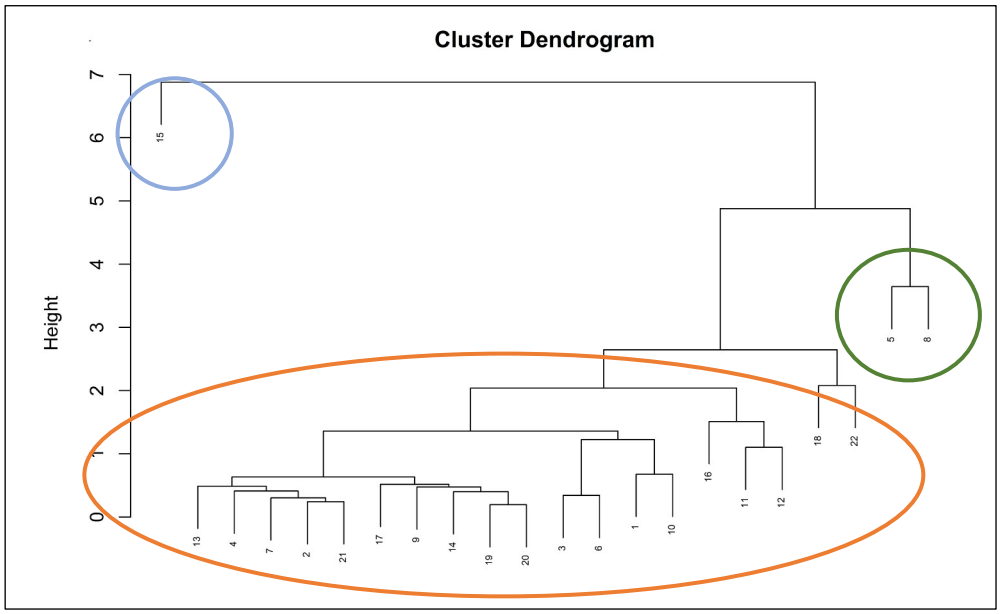


Figure 24: Dendrogram using Maximum Distance and Average-Linkage clustering method

The result obtained for the centroid and Average-linkage method is also fascinating, as it entirely agrees with the results obtained using the complete and McQuitty linkage methods. As shown in Figure 23 and 24, point 15, stands alone, points 5 and 8 forms a cluster, while the remain measurement points forms a larger group pf clusters with obviously distinct sub-clusters. The results obtained here essentially gives a great insight into how the measurements points should be grouped, as it further validates some of the results earlier obtained. Nevertheless, attempts were still made to further analyse the data, and determine the best way of grouping the various measurement points, through the use of Canberra and Minkowski methods of distance calculations.

3.7 Dendrograms Generated Using Canberra Distance Method

Canberra distance method involves the statistical measure of the distance between pairs of observations in a vector space (Lance & Williams, 1996; Giuseppe, Samantha, Roberto, & Cesare, 2009). For instance, the Canberra distance d between vectors p and q in an n-dimensional real vector space is given as:

$$d(p,g)=\sum_{i=1}^{n} \frac{\left| p_i - q_i \right|}{\left| p_i \right| + \left| q_i \right|}$$

Where;

p = $(p_1, p_2, p_3, p_4 \ldots\ldots\ldots p_n)$ q = $(q_1, q_2, q_3, q_4 \ldots\ldots\ldots q_n)$are vectors

Therefore, the method, together with all the clustering methods were implemented in this section in order to realise the probable numbers of cluster in the dataset.

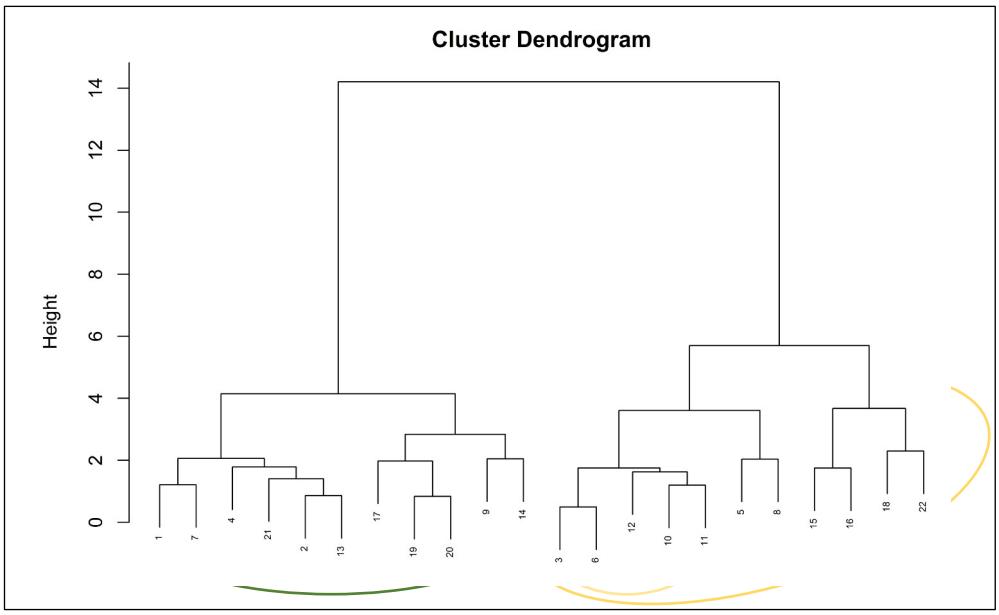


Figure 25: Dendrogram using Canberra Distance and Ward.D clustering method

Using Canberra distance and Ward.D clustering method, Figure 25 revealed that two major clusters can be obtained from the data sets. These two major clusters can be further broken down to obtain four distinct clusters as shown in Figure 25. Although, other sub-clusters are obviously seen to be present in the dendrogram.

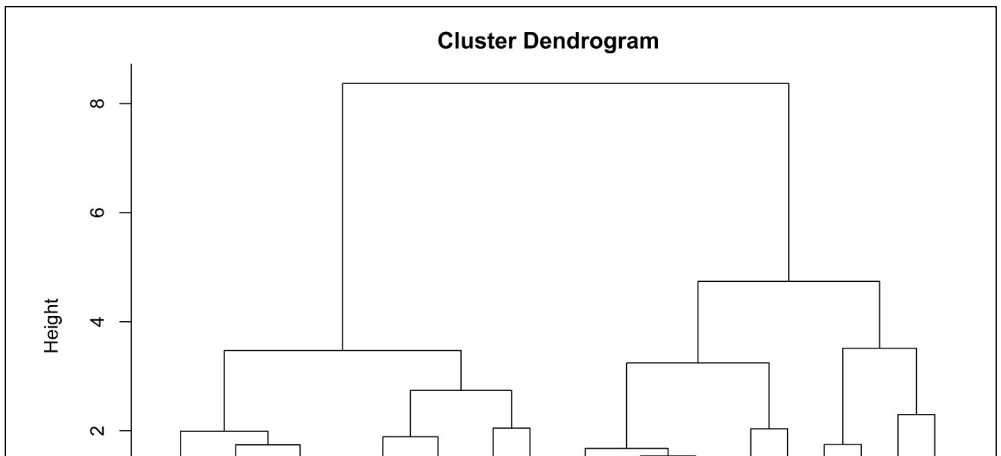


Figure 26: Dendrogram using Canberra Distance and Ward.D2 clustering method

The result here (in Figure 26) is entirely similar to that depicted in Figure 25, both in terms of numbers of clusters, and as well as the composition of elements in main clusters and as well as sub-clusters. There are two major clusters here, with probable four sub-clusters.

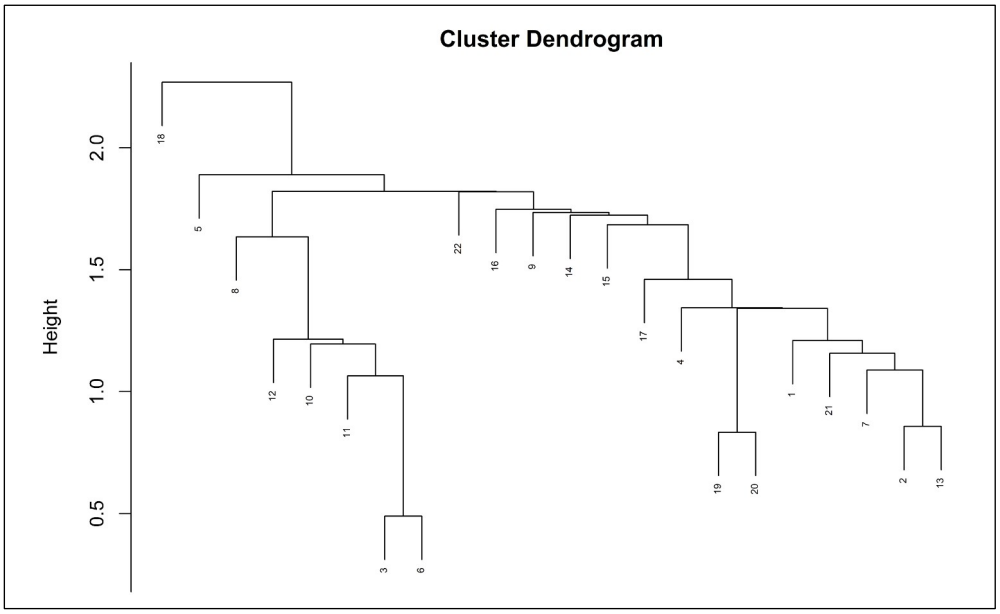


Figure 27: Dendrogram using Canberra Distance and Single-Linkage clustering method

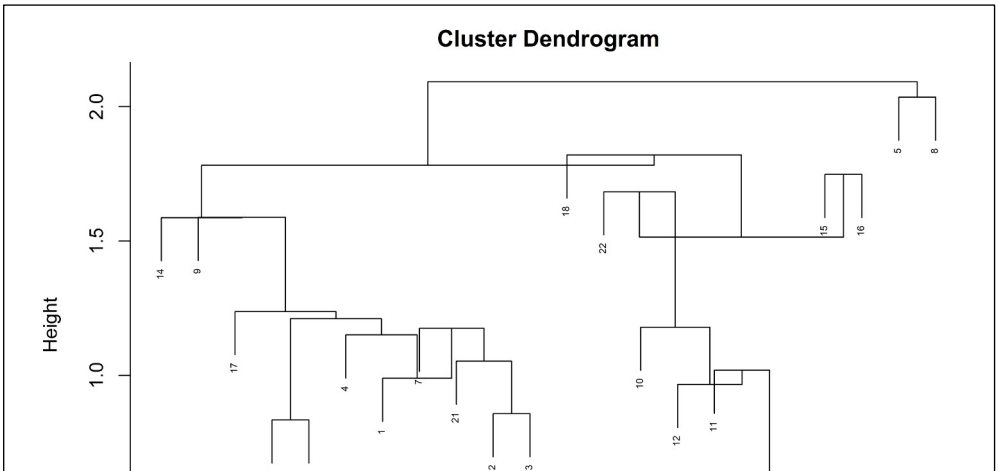


Figure 28: Dendrogram using Canberra Distance and Median clustering method

The results of the algorithm depicted in Figure 27 and 28 exhibits a rather complex groupings of the dataset, different from almost all the previous plots. It is quite difficult to discern the numbers of clusters present in the two dendrograms. However, worthy of note is the fact that, despite the clumsiness, in Figure 28, points 5 and 8 still forms a cluster.

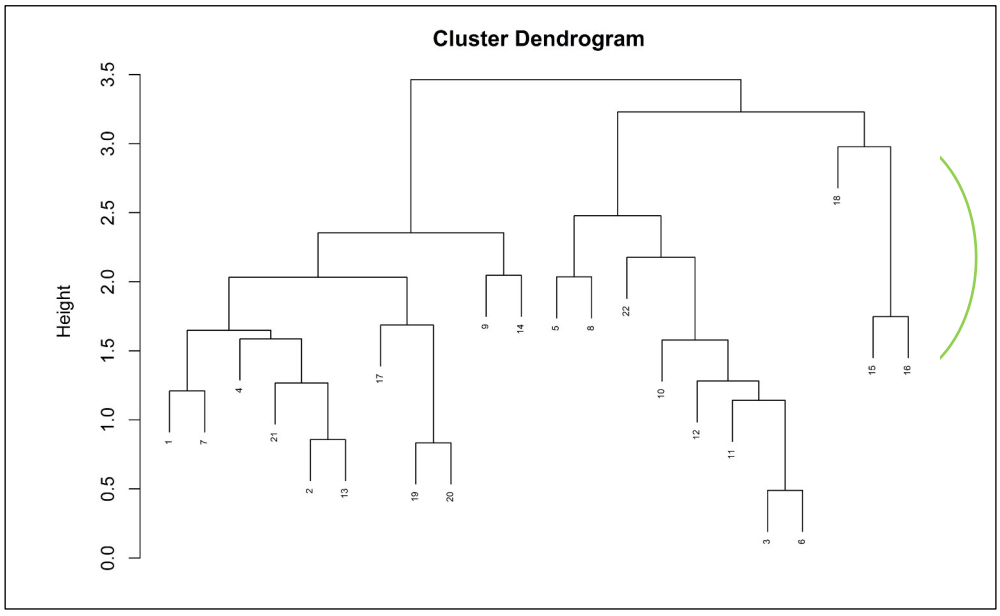


Figure 29: Dendrogram using Canberra Distance and McQuitty clustering method

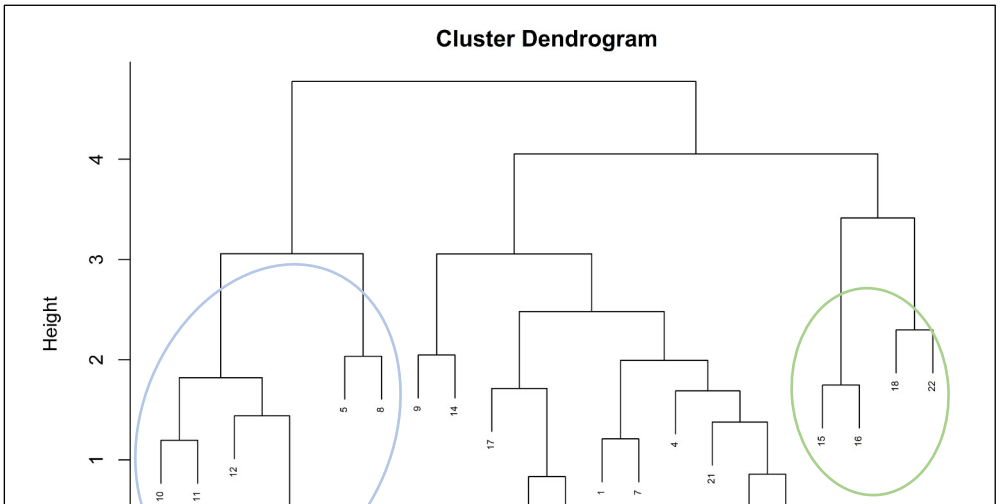


Figure 30: Dendrogram using Canberra Distance and Complete clustering method

The numbers of clusters present in Figures 29 and 30 seem obvious, unlike the two previous dendrograms depicted in Figure 27 and 28. From Figure 29 and 30, it is obvious that there are three groups of clusters, although with different compositions of measurement points within each of the clusters in the two dendrograms. There are also smaller groups of clusters within each of the main clusters.

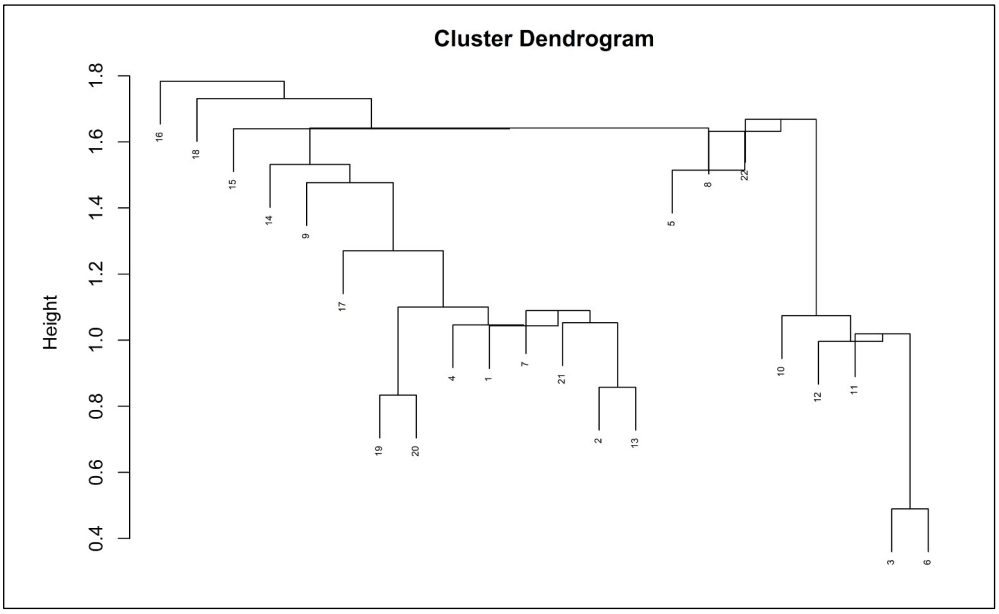


Figure 31: Dendrogram using Canberra Distance and Centroid clustering method

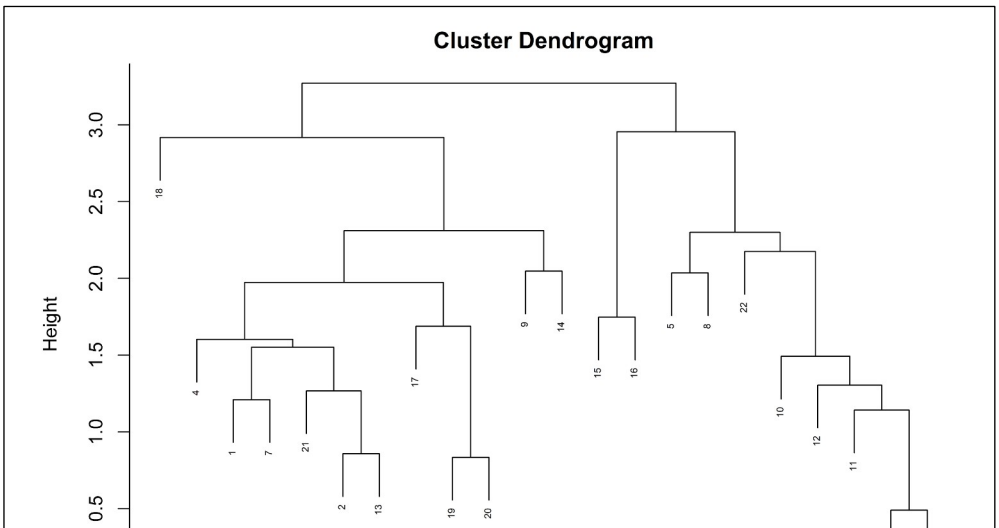


Figure 32: Dendrogram using Canberra Distance and Average clustering method

The algorithm implemented in Figure 31 and 32 also produces a somewhat complex result of clustering, in the sense that it is quite unclear the numbers of clusters present in the dendrograms. This is especially true for the dendrogram depicted in Figure 31. For Figure 32, it can be argued that there are three clusters, with many sub-clusters, out of which, measurement point 18, stands alone. Generally, in most cases, the Canberra method produces results that makes it challenging to discern the number of clusters present in the data set.

3.8 Dendrograms Generated Using Minkowski Distance Method

The last algorithm implemented in this section is Makowski distance method, which is basically the generalization of both the Euclidean distance and the Manhattan distance. It is defined is such a way that the Minkowski distance of order p (where p is an integrer) between two pooints;

$$X = (x_1, x_2, x_3, x_4 \ldots\ldots\ldots x_n)\ Y = (y_1, y_2, y_3, y_4 \ldots\ldots\ldots y_n) \varepsilon\ R^n$$

Is defined s;

$$d(X, Y) = ¿\ |\text{^}2]\ \text{^}(1/p)$$

This distance method, with other clustering method earlier discussed is therefore implemented in this section, and the various dendrograms that were obtained are therefore depicted in subsequent pages.

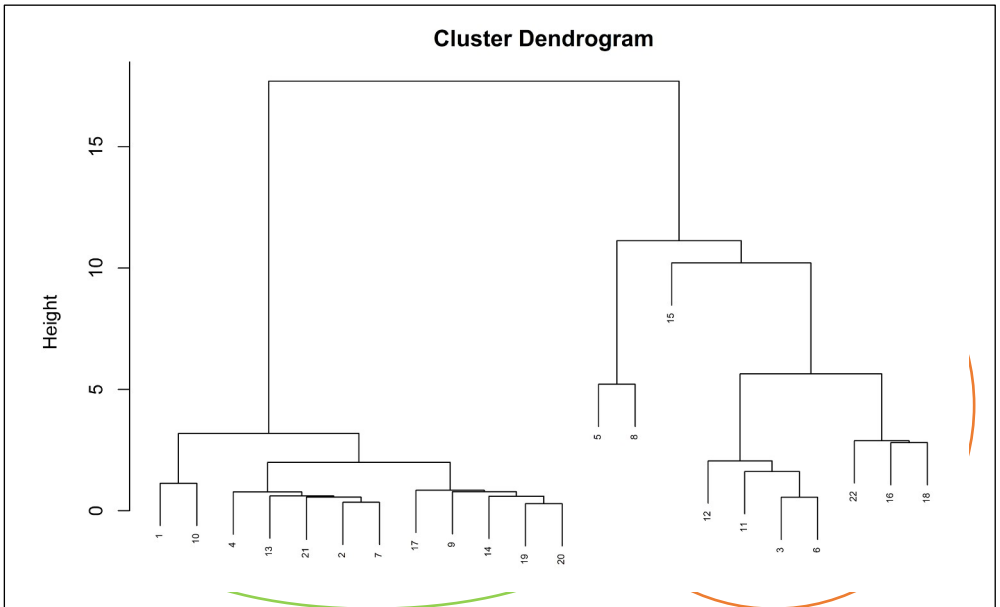


Figure 33: Dendrogram using Minkowski Distance and Ward.D clustering method

The first attempt in this section involves the combination of Minkowski distance with Ward.D clustering method. From figure 33, it is obvious that that are basically two clusters in the dataset. Although it can be argued that there is a third major cluster which is formed by the agglomeration of measurement 5 and 8. Within the first main cluster, there are three sub-clusters, and within the second main cluster, there are about three sub-clusters.

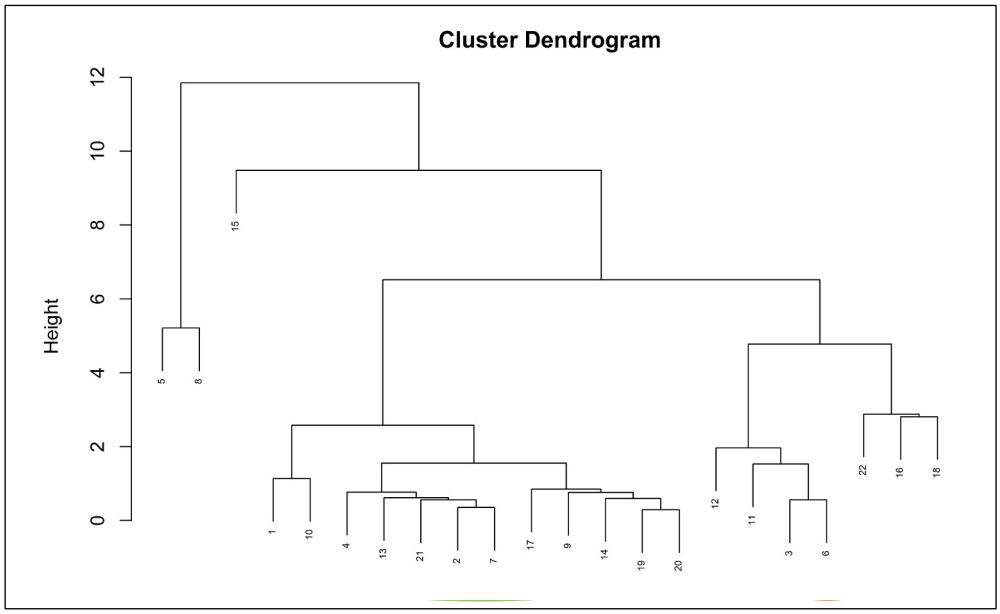


Figure 34: Dendrogram using Minkowski Distance and Ward.D2 clustering method

There appears to be four clusters in the dendrogram depicted in Figure 34. Again, point 5 and 8 forms a cluster. Point 15 stand alone, point 1, 10, 4, 13, 21, 2, 7, 17, 9, 14, 19 and20 forms another big clusters, while points 12, 11, 3, 6, 22, 16 and 18 forms another cluster.

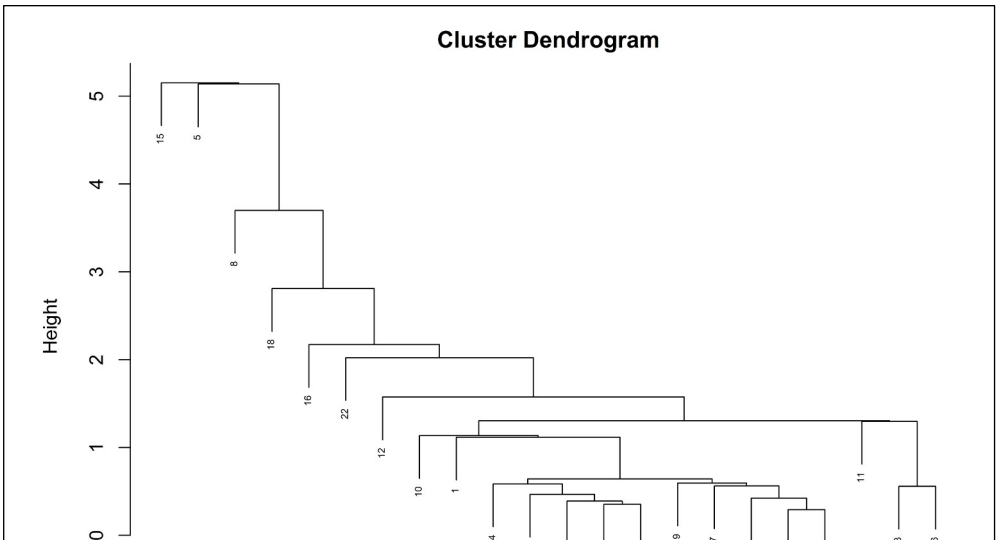


Figure 35: Dendrogram using Minkowski Distance and Single-linkage clustering method

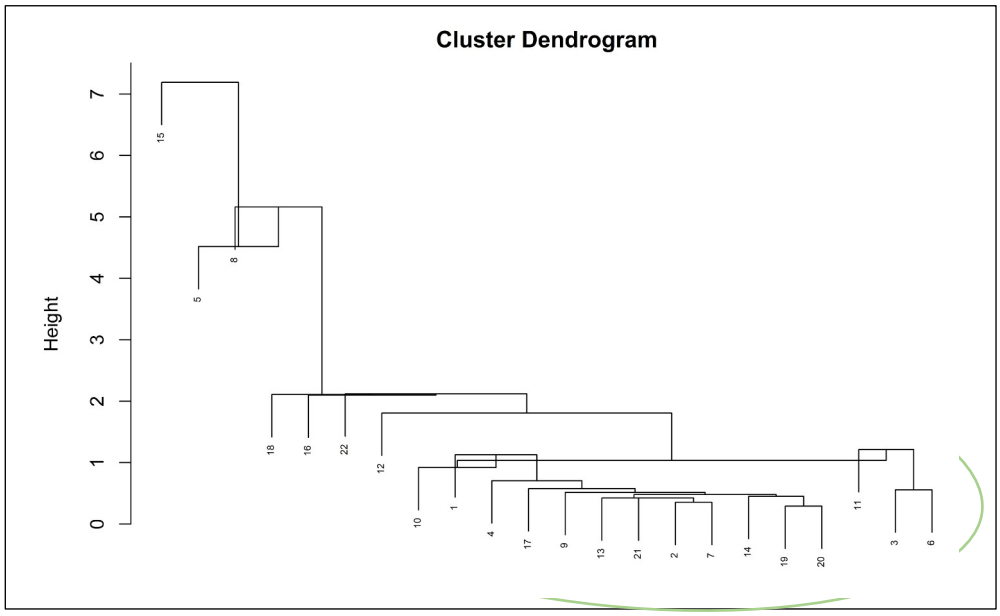
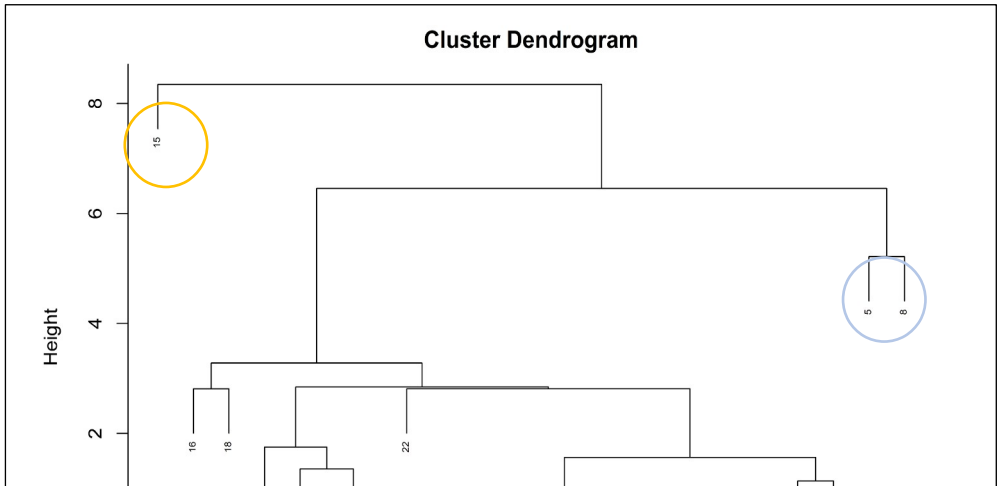


Figure 36: Dendrogram using Minkowski Distance and Median clustering method

Generally, the Single-linkage clustering method exhibits a somewhat vague grouping of the clusters, for all the methods of distances employed so far. This is because the single-linkage method approaches the distance between two clusters as the minimum distance between their members, it tends to produce long thin clusters in which nearby elements of the same cluster have small distances between them.

For dendrogram depicted in Figure 36, which is a result of the combination of Minkowski distance and Median clustering methods, there appears to be three main clusters in the dendrogram. The first include only measurement point 15, the second cluster includes measurement points 5 and 8, while the other measurement points arguably forms a larger cluster, with sub-clusters.



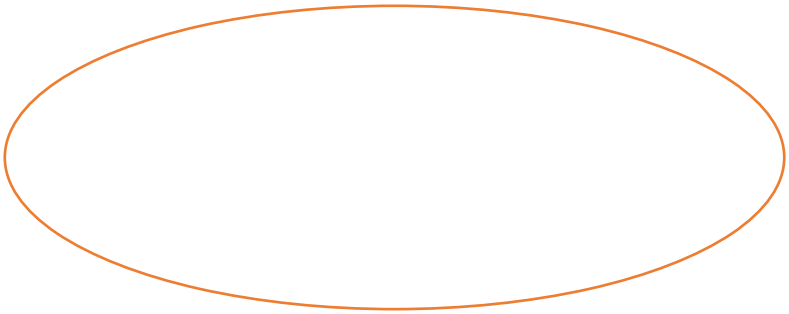


Figure 37: Dendrogram using Minkowski Distance and McQuitty clustering method

The dendrogram depicted in Figure 37 revealed that using the Mcquitty clustering method with Minkowski distance algorithm, the dataset could be segregated into three main clusters (though probable). Points 5 and 8 forms a cluster, points 15 stands alone, while the other points agglomerate together, though visibly divisible into two clusters as shown in Figure 37.

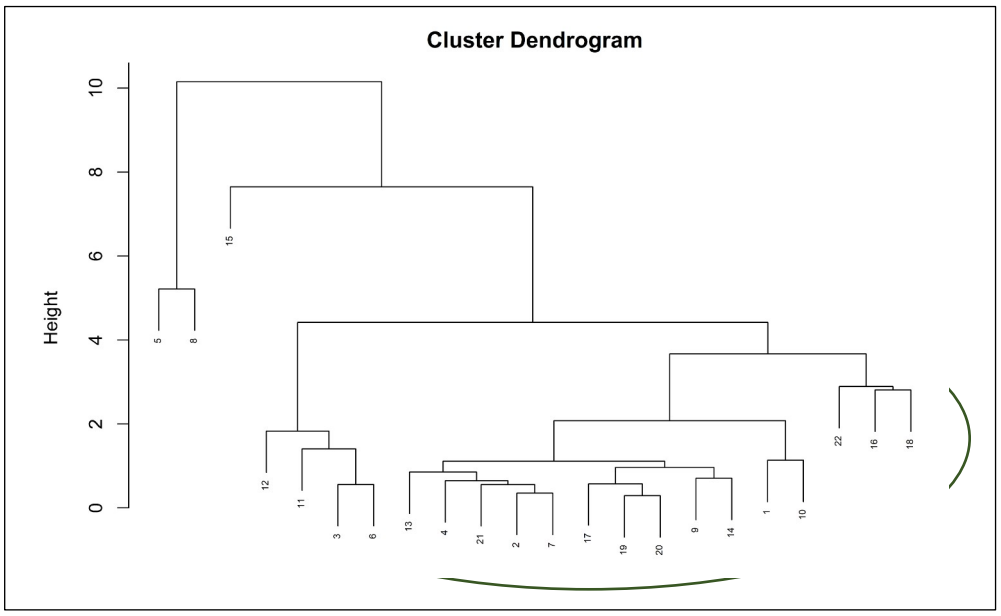


Figure 38: Dendrogram using Minkowski Distance and Complete clustering method

The dendrogram obtained using the complete clustering method is similar to that obtained with McQuitty method. As shown in Figure 38, points 5 and 8 combine to form a cluster, point 15 stands alone, and other points appears to be grouped together; though there are still separate clusters that are still evident within the group.

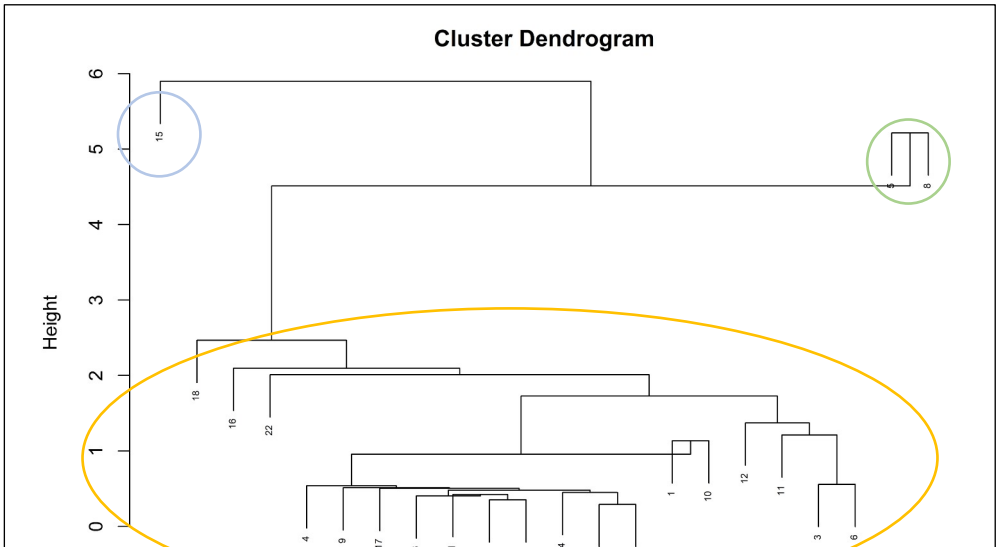


Figure 39: Dendrogram using Minkowski Distance and Centroid clustering method

The dendrogram depicted in Figure 39 revealed that there data sets comprise of three clusters. As usual, measurement point 15 stands alone, measurement points 5 and 8 stand forms a cluster, while the bulk of the remaining measurement points forms a larger cluster. Although, this larger cluster comprises of about sub-clusters, each of which can still be further treated as a sperate entity.

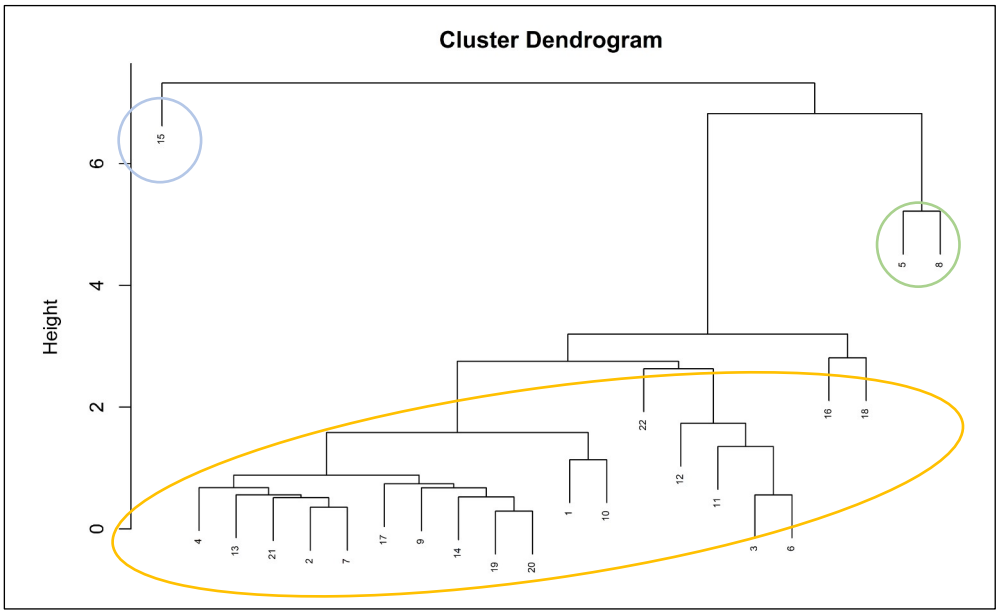


Figure 40: Dendrogram using Minkowski Distance and Average clustering method

Like the dendrogram depicted in Figure 39, the dendrogram depicted in Figure 40 shows that the dataset comprise of three clusters. Where measurement point 15 also stand alone, and measurement points 5 and 8 forms a separate cluster. There exists a larger agglomeration of points which consist of about 5 distinct clusters. The result of grouping obtained with Minkowski distance is entirely similar to that obtained for Euclidean distance method, perhaps, this is due to the fact that the Minkowski method is simply the simplification of both the Euclidean distance and the Manhattan distance methods.

The prevailing number of clusters in the various dendrograms analysed is three. To a large extent, majority of the dendrograms produce similar results both in terms of numbers of clusters and as well as composition of elements (that is measurement points) contained in each of the clusters. In most cases, measurement point 15 stands alone, while measurement points 5 and 8 often converge to form a distinct cluster. The remaining measurement points are often merged to form a larger cluster, though in most cases, with marked sub-clusters that

can be further sub-divided. In most cases, within this larger group, measurement points 2,4, 7, 9, 13, 17, and 21 are often together, while also measurement point 1 and 10 are always clustered together. Similarly, measurement points 3, 6, 11, and 12 are always grouped together, points 16, 18, and 22 are often together, while 14, 19 and 20 appears to always cluster together. It can be inferred from the various dendrogram plots generated in this analyses that measurement point 15, 5 and 8 exhibits some heterogeneity, which often sets them further apart from other measurement points. There is arguably some homogeneity in the remaining measurement points, hence they often converge.

3.9 Selection of Preferred Clusters, and Dendrogram with Justification

Based on the various results analysed, the dataset considered in this analysis will be broadly grouped into three clusters. This to a great extent reflects the numbers of species of penguins that toll the study area. The most repeated form of grouping in the various algorithms includes a cluster formed by 5 and 8, another cluster formed by 15 alone and the third cluster formed by the remaining data sets. As shown in Table 3 and Figure 41, this pattern was obtained in more than 50% of the dendrograms generated in this analysis. The algorithms that produced this preferred and most prevalent grouping of the dataset considered in this analyses are given in the Table below;

Table 6: Various Algorithms combination that produces preferred grouping of the dataset.

Algorithm (Method), and Figure	Cluster Composition	No of Custers
Figure 4: Euclidean Distance and Median clustering method	5 & 8; 15; and Others	3
Figure 5: Euclidean Distance and McQuitty clustering method	5 & 8; 15; and Others	3
Figure 6: Euclidean Distance and Complete clustering method	5 & 8; 15; and Others	3
Figure 7: Euclidean Distance and Centroid clustering method	5 & 8; 15; and Others	3
Figure 8: Euclidean Distance and Average clustering method	5 & 8; 15; and Others	3
Figure 11:: Manhattan Distance and Single-Linkage clustering method	5 & 8; 15; and Others	3
Figure 13: Manhattan Distance and McQuitty clustering method	5 & 8; 15; and Others	3
Figure 14: Manhattan Distance and Complete-Linkage clustering method	5 & 8; 15; and Others	3
Figure 15: Manhattan Distance and Centroid clustering method	5 & 8; 15; and Others	3
Figure 16: Manhattan Distance and Average-Linkage clustering method	5 & 8; 15; and Others	3
Figure 19: Maximum Distance and Single-Linkage clustering method	5 & 8; 15; and Others	3
Figure 20:: Maximum Distance and Median clustering method	5 & 8; 15; and Others	3
Figure 21: Maximum Distance and McQuitty clustering method	5 & 8; 15; and Others	3
Figure 22: Maximum Distance and Complete-Linkage clustering method	5 & 8; 15; and Others	3
Figure 23: Maximum Distance and Centroid clustering method	5 & 8; 15; and Others	3
Figure 24: Maximum Distance and Average-Linkage clustering method	5 & 8; 15; and Others	3
Figure 36: Minkowski Distance and Median clustering method	5 & 8; 15; and Others	3
Figure 37: Minkowski Distance and McQuitty clustering method	5 & 8; 15; and Others	3
Figure 38: Minkowski Distance and Complete clustering method	5 & 8; 15; and Others	3
Figure 39: Minkowski Distance and Centroid clustering method	5 & 8; 15; and Others	3
Figure 40: Minkowski Distance and Average clustering method	5 & 8; 15; and Others	3

Source: Rowland's Analysis, 2021

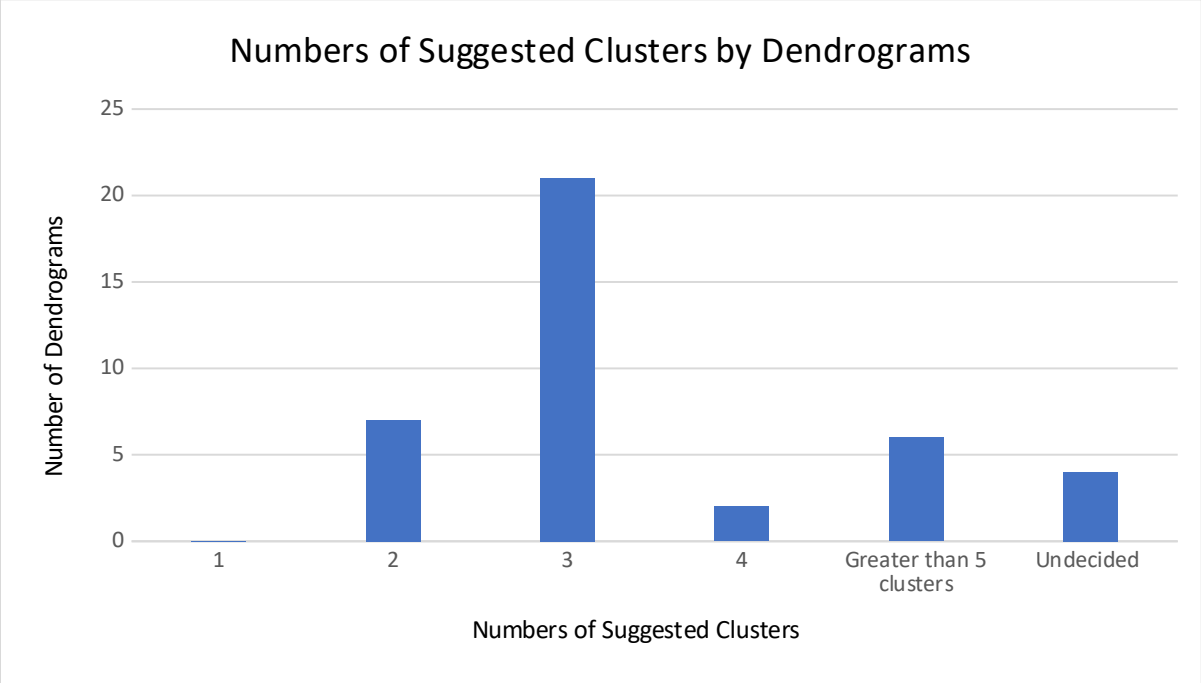


Figure 41: Number of Suggested Clusters by Dendrograms

However, a final decision was made to select the result produced by the combination of Euclidean distance, and average-linkage clustering algorithms depicted in Figure 8 (Shown in Figure 42 below). This was deemed appropriate as Euclidean distance method is the most widely accepted distance algorithm due to its reliability and stability (Dibya et al, 2014). When combined with average-linkage method, as the case here, the algorithm can produce an incredibly accurate result, because the average linkage defines the distance between two clusters as the average pairwise distance between genes in cluster, which helps to improve reliability of results, and consequently eliminate bias (Mohammad & Zhong-Hui,, 2015).

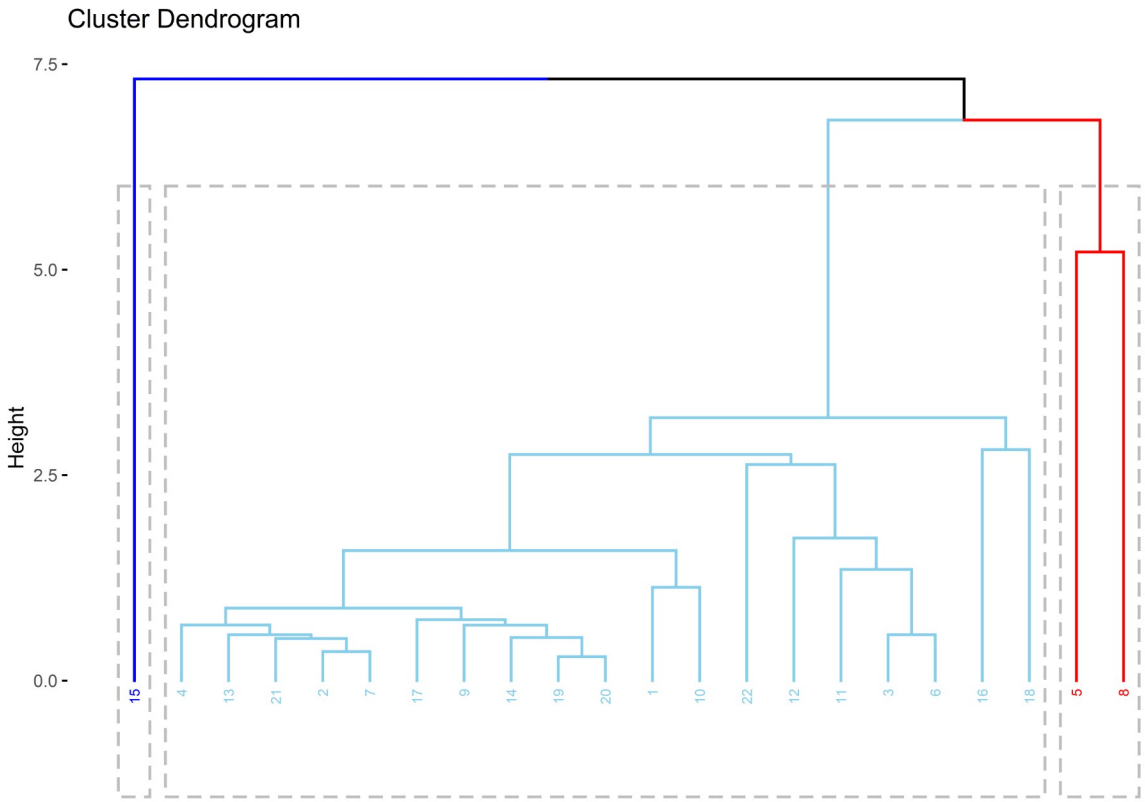


Figure 42: Euclidean Distance and Average clustering method (Plotted with factoextra)

3.9.1 Properties of Selected dendrogram (Cluster adopted with the Measurement Points)

From Figure 42, there are three main clusters in the dendrogram produced by the preferred algorithm. The first cluster includes measurement point 15, the second clusters includes measurements points 4, 13, 21, 2, 7, 17, 9, 14, 19, 20, 1, 10, 22, 12, 11, 3, 6, 16 and 19. While the third cluster includes measurement points 5 and 8. The second cluster, however, comprised of sub-clusters; Within the second main cluster, measurement points 4, 13, 21, 2 , 17, 9, 16, 19 and 20 belongs to a sub-cluster. Measurement points 1 and 10 belongs to another cluster, while the remaining points can be said to belong to another separate sub-clusters, and can still be further sub-divided.

3.9.2 Verification of the Selected Number of Cluster

This section essentially seeks to prove whether the number of selected clusters is reasonable enough for the dataset considered in this analyses. Two of the popular ways to determine this, is through the use the elbow plot and the silhouette method provided by the kmeans algorithms. The elbow plot specifies that the ideal numbers of clusters is the point where the dataset stops changing rapidly. Whereas, the silhouette plot displays a measure of how close each point in each cluster is to points in the neighbouring clusters and thus provides a way to visually assess the number of clusters present in a dataset.

The elbow plot was achieved with the aid of the following code;

```
fviz_nbclust(data, kmeans, method='wss')
# And the one below, after it was obvious the function stops decreasing rapidly at point 3
+ geom_vline(xintercept=3, linetype=2)
```

As depicted in Figure 43, the data changes rapidly up till around point three, where the change began to be gradual onwards.

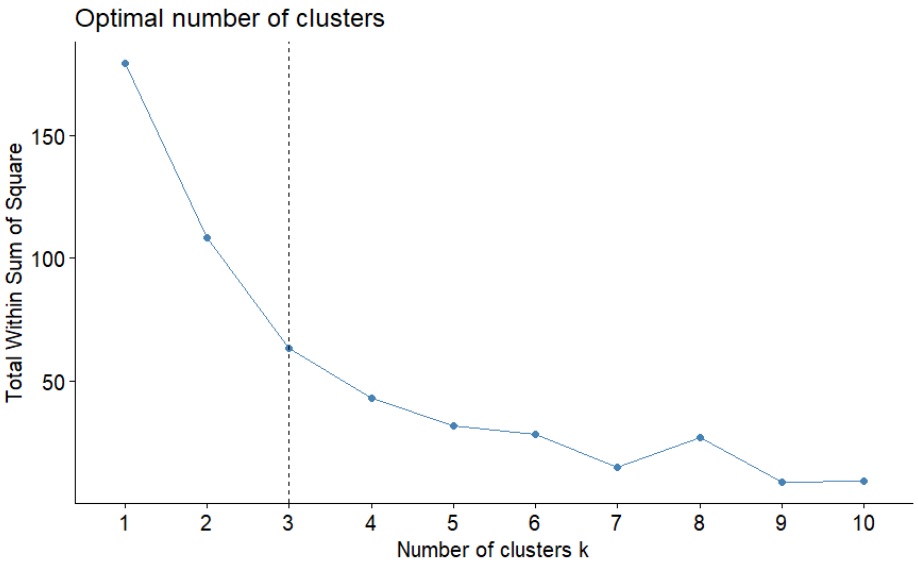


Figure 43: Elbow Plot for Validating Optimum number of clusters.

Further verification was attempted with the aid of the silhouette method using the code given below;

```
fviz_nbclust(data, kmeans, method = "silhouette")
```

However, the method produces a different result. As depicted in Figure 44, it suggests that the optimal number of clusters in the data set is 2, as two has the highest score on the plot, and closely followed by 3. Meanwhile, researchers suggest that two clusters is never an ideal solution for grouping a data set (that is considerable large, but not too small) into clusters, because in most cases, there is a natural tendency that all datasets will have two clusters (at least). Therefore in this analyses, a decision was finally made keep the the dataset in three clusters, as earlier determined by the dendrogram depicted in Figure 42.

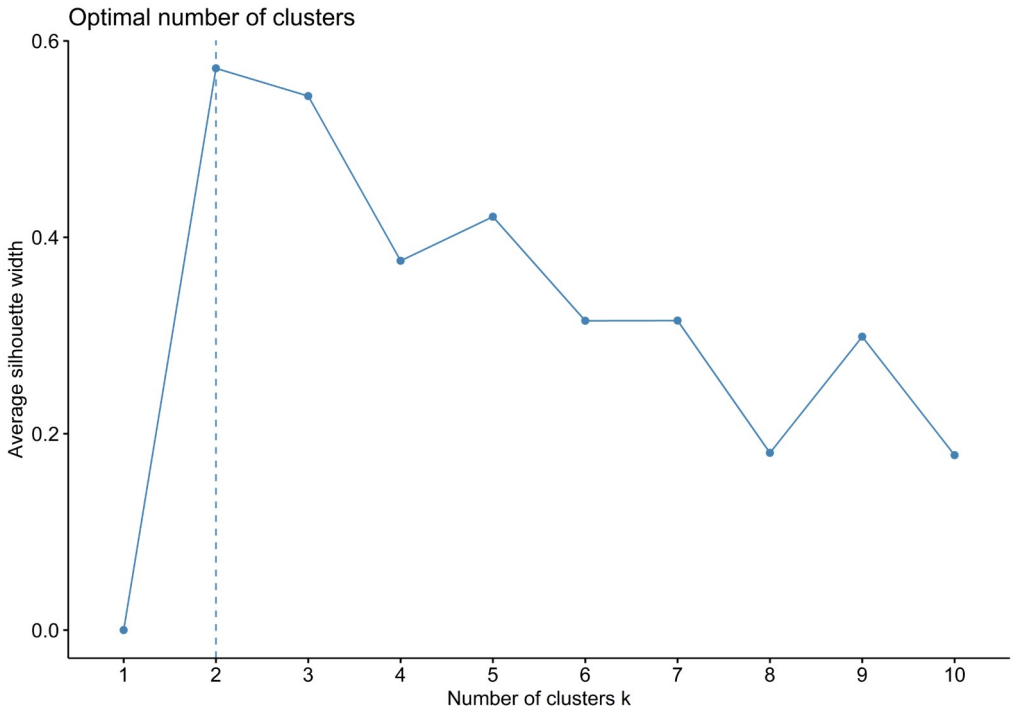


Figure 44: Silhouette Plot for Validating Optimum number of clusters.

4.0 Conclusion

The various analyses conducted have revealed both heterogeneity and homogeneity among the different measurement points, based on the nutrient data recorded at these locations. These results allow for a rational and logical decision regarding which measurement points indicate a similar impact of penguins in the study area. Across most of the algorithms used, measurement point 15 is often set apart from the others. Similarly, measurement points 5 and 8 exhibited some similarities and frequently clustered together, while the remaining measurement points tended to cluster together, sometimes forming distinct sub-groups.

Based on this analysis, it can be concluded that, apart from measurement points 5, 8, and 15, the penguins affect the other areas at a relatively similar rate, though with some notable differences among the points. For example, points 4, 13, 21, 2, and 7 were affected at more similar rates, as were points 17, 9, 14, 19, and 20; points 1 and 10 clustered together; points 22, 12, 11, 3, and 6 were also grouped similarly; and points 16 and 18 were affected at a similar rate. These results are consistent with the findings obtained from the three reduced matrices in Excel, where points 2 and 7 clustered together, as did points 19 and 20.

In contrast, points 5 and 8 are affected at rates that are uniquely different from other areas, while point 15 is affected at a rate that is significantly different from both points 5 and 8 and from the other points mentioned earlier. It may also be inferred that the impact of the three species of penguins on the soil and surface waters will be greater in areas that converge to form larger clusters.

5.0 References

Bradley, B. (2021). *Hierarchical Clustering.* Retrieved from https://bradleyboehmke.github.io: https://bradleyboehmke.github.io/HOML/hierarchical.html

Carnegie Mellon University. (2009). *Distances between Clustering, Hierarchical.* Retrieved from https://www.stat.cmu.edu: https://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf

Clemens, J. (2019,). *Introduction to Hierarchical Clustering.* Retrieved from https://towardsdatascience.com: https://towardsdatascience.com/introduction-hierarchical-clustering-d3066c6b560e

Dibya, J., & et al. (2014). Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab . *Journal of Computer Science and Information Technologies*, 2501-2506.

Giuseppe, J., Samantha, R., Roberto, V., & Cesare, F. (2009). Canberra Distance on Ranked Lists", in Shivani Agrawal. *Advances in Ranking* (pp. 22–27). NIPS 09 Workshop.

James, G. (2015). Validating a Hierarchical Cluster Analysis. https://www.youtube.com/watch?v=mSzk2KrbNfs.

Krish, N. (2019). Hierarchical Clustering Intuition. https://www.youtube.com/watch?v=0jPGHniVVNc.

Lance, N., & Williams, W. (1996). Computer programs for hierarchical polythetic classification (similarity analysis). *Computer Journal*, 9 (1): 60–64.

Loga, M. (2021). *Project Lecture Material .* Warsaw: Warsaw University of Technology: Unpublished .

Mohammad, S., & Z.-H. D. (2015). Hierarchical k-Means. *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology*.

University of Alberta. (Not Available). *Multivariate Fundamentals: Distance.* Retrieved from https://sites.ualberta.ca: https://sites.ualberta.ca/~lkgray/uploads/7/3/6/2/7362679/slides-clusteranalysis.pdf

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58, 236–244.