Assignment for Multivariate Statistical Analysis - WiSe 2020/21

Dendrogram Computations & Interpretations

Name: Mert Erdinc

Matriculation Number: 615804

Study Subject: Master of Science in Business Administration

Place & Date: Turkey 19/03/2021

R is used as a statistical software program in this report.

1. Clustering

Clustering is an unsupervised multivariate data analysis technique that essentially aims to group the data points in a given data set that are similar or close to each other and uncover the hidden clusters so that useful insights could be derived from them depending on the aim of the analysis. Essentially, clustering tries to achieve two objectives at the same time by trying to simultaneously maximize the homogeneity within the groups and the heterogeneity between the groups and nowadays it has a variety of areas of use regarding real-world applications such as Biology, Business, Social Sciences, Computer science and many more. While a specific clustering method may be known to produce better results with a certain area of application or a data set, to get the most desired result, one usually needs to iterate through different clustering methods like hierarchical clustering, k-means, model-based clustering, density-based clustering. In most of the cases, even some clustering decisions like whether Mahalanobis or Euclidean distance will be used for the calculation of the distance matrix or what will be the number of clusters to be formed, which, unlike hierarchical clustering, needs to be specified a priori in clustering methods like k-means or model-based clustering needs to be tried to decide on a good enough combination of "parameters" for the clustering. Also, for different clustering methods, to visualize the clustering results (in the case of hierarchical clustering, the whole process) different tools are used; namely, in the case of K-means the distribution of the data points and how they are clustered could be effectively shown using Multi-dimensional Scaling Method (MDS) by plotting a lower dimensional version of the clustered data set, and in the case of model-based clustering the results are mainly visualized by plotting the uncertainty of the data points or the density of the clusters. So, including hierarchical Clustering, the results of each clustering method is presented using different visualization techniques and different kinds of fit values to compare the performance of the clustering decisions such as total sum of squares, elbow curve, BIC value and so on.

In the case of hierarchical Clustering, because of the approach it uses for clustering, dendrograms are used, which will be the focus of this report; however, it is also useful to keep in mind that, if wanted, dendrograms could be supported by other multivariate visualization techniques such as parallel coordinate charts, where the effect of the chosen variables for the clustering on the heterogeneity of the data points in each cluster could be observed and visualized. Before proceeding with the next section, providing a brief introduction of the idea of hierarchical clustering is useful as it will not only serve as a basis for understanding it but also will make the interpretation easier. Hierarchical clustering groups the data points by building a tree of clusters and what one may call "sub-clusters", where a new cluster is formed by merging the cluster from the last step and a new data point or a cluster together based on different techniques such as Ward, Linkage or Centroid methods while doing that. If the clustering uses a "bottom-up" approach, where it starts clustering with the assumption that at first each data point constitutes a cluster by itself and then clusters them until one big cluster is achieved, it is called agglomerative hierarchical clustering, which is further divided into different approaches for clustering such as the ones that are mentioned just above. If the clustering uses a "top-down" approach, where it starts clustering with the assumption that each data point belongs to one big cluster and then divide them until every one of them becomes a cluster of itself, it is called divisive hierarchical clustering. The focus in the next sections will be on how dendrograms are useful, how

they are used in the interpretation of the clustering results while at the same time topics like how each clustering method (Linkage, Variance and Centroid) is calculated, what advantages and disadvantages they might have over each other and lastly comparison of different clustering methods applied to the given data set and the dendrograms they produce will be examined.

2. Dendrograms

As mentioned in the previous section, agglomerative hierarchical clustering method starts clustering assuming that every observation is a cluster itself and forms bigger clusters sequentially by merging clusters according to some criteria aiming to both increase the homogeneity within the clusters and the heterogeneity in between them. This approach is sometimes also called as "bottom-up" approach, which is used more often and tends to produce better results than the divisive hierarchical clustering. Divisive hierarchical clustering, on the other hand, uses the exact same approach and process that the agglomerative hierarchical clustering uses except that it starts forming clusters by dividing one big cluster into sub-clusters until every data point becomes a cluster by itself. What dendrograms essentially do is that to visualize how the process of hierarchical clustering looks like depending on the method used, how similar or different some points and clusters compared to each other are and at which specific value are the points or clusters joined to another cluster (height values). Height values are the values at which the decision of merging two specific clusters is made. For example, if the clustering criteria is the minimum distances between clusters and the minimum distance at an example is 2 in the distance matrix between two specific clusters among all the other combinations of clustering possibilities, the point where these two clusters are merged matches with a value of 2 in the y-axis, which is also called "height". Also, dendrograms have one root node and are divided further into binary leaf nodes in every re-clustering step until there is no more step of clustering is left possible. For simplicity, one can imagine dendrograms as continuous decision trees, where if the distance of a data point to the closest cluster is the smallest one among the others (Linkage Methods) or the inclusion of a data point results in the minimum variance within a cluster (Ward), the data point is included in the cluster. If the data point doesn't fulfil criteria like above, they will be eventually included to a cluster either through the end or the beginning of the dendrogram as there is no left-out data points in dendrograms. An efficient dendrogram usually clusters most of the data points early in the height values and also is close to be symmetrical, both of which could be interpreted as signs of distinct cluster formation but there could be cases where a data point is included to a cluster at a very late stage of the clustering (high height value). This could be either because of the method used, for example Single-Linkage in which such occurrences are typical, or because that data point is an outlier, both of which will be examined in more detail in the next parts of the report.

One other important aspect of the dendrograms to consider is that it is generally not realistic to decide on which clustering method and parameters produce the best result by solely relying on the dendrogram itself. It is, however, useful to see which methods are good candidates that needs more examination and which ones are bad ones that can be easily eliminated from further consideration. For instance, it might depend on the given data set and the desired output too, but a decision one might directly make is not to consider single-linkage method for further examination because of its usual chain-like dendrogram output and similarly one can directly try other clustering decisions on the Ward method as it tends to produce good results compared to other methods. Another benefit that

dendrograms provide is that it allows researchers to be able to roughly see the distribution of the data points in the clusters depending on the decision of number of clusters. Researchers usually want to see clusters that all include at least some proportion of the observations in each cluster so that clusters are meaningful to use rather than totally unbalanced proportions of observations in each cluster such as 2%, 3%, 5% and 90%. Also, as briefly mentioned above, although the more data points there are in the analysis, the harder it gets to comment on them individually, one might observe some data points that are included in a cluster shortly before the end of the clustering process. This can be useful because, with further analysis of these points, one can draw conclusions about whether these points stick out in the dendrogram because they are outliers or because of the specific clustering method used, which can play an important role in the analysis. Also, because most of the time there will be a substantial amount of data in the analysis thus also in the dendrogram, individual comparison of the similarities of the data points won't be looked into and usually is not necessary. Because of the same reason, dendrograms serves as an explanatory data analysis tool that shows the way that the clustering process is carried out and as well as that provides a first look at how the hierarchical clustering is done.

3. Dendrograms explained with examples

In this section a more detailed explanation on dendrograms will be provided supported with some examples. There are some important aspects and points about dendrograms which could be crucial to know before using them. One issue that needs to be addressed regarding dendrograms is the fact that in hierarchical clustering one chooses the desired number of clusters after the clustering is done. However, contrary to common belief, dendrograms do not provide a certain number of clusters to produce an "optimal" result but rather they only provide some insights that one can take into account to make a clustering decision while also considering other things like which variables will be used and the context of the data, which is not provided in the given data set as it is randomly produced. One other fact that is important to consider is that dendrograms are more accurate in shallow levels rather than high levels but this will be discussed in more detail below.

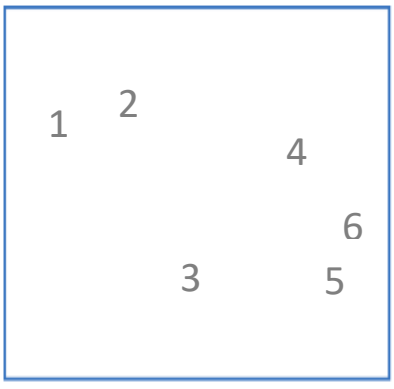

Figure 1. Plot of the data points

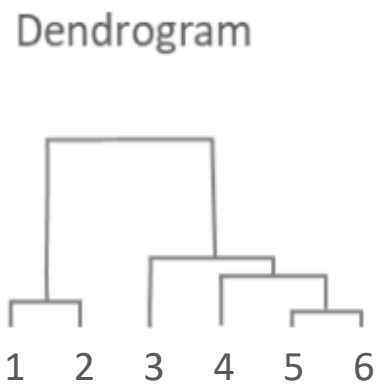

Figure 2. Example Dendrogram

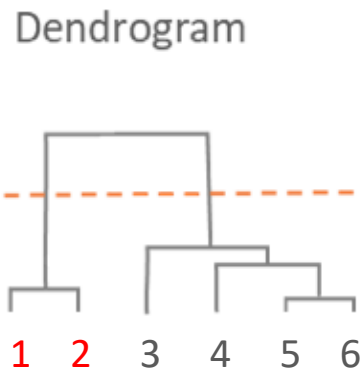

Figure 3. Example partitioned Dendrogram

In figure 1, 2 and 3 a MDS plot and two imaginary dendrograms are presented respectively to illustrate how a dendrogram looks like and to further elaborate on issues regarding how it can be interpreted. Unlike most of the real-world applications, these figures are kept basic with the intent of having higher ability of interpretability and demonstration.

Hierarchical clustering and thus dendrograms are created using some clustering criteria such as minimization/maximization of distance in between data points, a mix of the two, variance-based methods, centroids and a couple more techniques in addition to these. Each clustering decision is done in a way that maximizes the homogeneity within the new cluster and the heterogeneity between the clusters so that they are as distinct as possible. The y-axis, in other words height, refer to values on which the clustering process is based on and they are generated using different hierarchical clustering methods. After each merge of clusters, the calculations are repeated with the new set of data that includes the merged clusters from the last iteration and a new cluster is formed until a new one cannot be formed anymore. As stated earlier, with huge data sets individual examination of the data points usually becomes insignificant to look at but in cases where a smaller data set is analysed, one can conclude that points that are grouped first (when the height is relatively low) are the ones that are most similar to each other and the clusters that are clustered relatively late in the process, where the height is relatively high, are dissimilar from each other. Both of these insights can be confirmed by the MDS plot in figure 1. However, one important aspect to pay attention to is that there will be always a loss of information in the dendrograms unless the ultrametric tree inequality is satisfied, which is not very likely to be the case in real-world data sets. Last thing that one needs to be careful about is that due to the nature of dendrograms the interpretation in the similarity of the data points gets less accurate as the height increases. In the above figures, at hind-sight one might think 3 and 4 are much more similar to each other than 3 and 2 are but, in reality, in Figure 1 one can observe that the difference is not very large. Also, one more thing that one needs to be careful about when interpreting the dendrogram is that one shouldn't conclude that 3 and 4 are similar to each other. They are next to each other in the dendrogram because their relative distance to 4,5,6 cluster is similar to each other. Lastly, on figure 3 how could a clustering decision be made is demonstrated with 2 clusters. From the given tree cut, one can observe two clusters, one composed of numbers 1 and 2 marked with red and the other composed of numbers 3, 4, 5 and 6 marked with the colour black.

4. Calculation of Hierarchical Clustering Methods and the Dendrograms

In this section the way each hierarchical clustering method is calculated, what kind of advantages and disadvantages they have on each other and the kind of dendrograms they produce will be examined.

As mentioned in the previous sections, hierarchical clustering methods are divided into two as Agglomerative and Divisive. It might depend on the data set and the purpose of the clustering but for most of the time Agglomerative Hierarchical Clustering is used, which is further divided into some subgroups. In this report 3 of them will be examined, which are namely Linkage methods where clustering is done based on the distances between all objects, Variance method where clustering is done in a way that tries to minimize the variance within the clusters and Centroid method where the clustering is done in a way that minimizes the distances between clusters' means. Also, there are many decisions one could take that can influence the shape of the dendrogram. Some examples of these decisions are that which scaling method is used, whether the Euclidean, squared Euclidean, Mahalanobis or (in cases where the data is mixed, i.e. binomial, nominal, continuous etc.) Gower metric distance matrix will be used and which variables to be used in the calculation of the distance matrix. Because iterating through these decisions would only point out the effect they have on clustering and wouldn't specifically contribute to interpretation of the dendrograms, they have been held constant through all clustering methods. Thus, for scaling, z-scores of each value are found, for

the distance matrix calculation Euclidean metric is used and all of the 12 variables were included as there were no context in the data set. So, in the corresponding sections below, each hierarchical clustering method is calculated, advantages and disadvantages they have and the interpretation of the dendrograms they produce will be discussed.

4.1 Single Linkage Method

In the single linkage method, the clustering process is done based on the minimum distances between clusters or data points. To be more specific, what single linkage algorithm does is that every time before it merges clusters it calculates a new distance matrix including the newly merged clusters from the last iteration. The elements of this matrix consist of the minimum distance either between points of two clusters or a data point and points of a cluster. Based on it, the objects who have the minimum distances between each other is merged or combined into a cluster and this process is repeated with the updated distance matrix until all of the data points are put in one big cluster. Below, the formula numbered 1, shows the formal expression of the single linkage algorithm's objective function and the idea behind it, where the minimum distance between two clusters, namely R and S is defined by the minimum distance between its members, a_i and a_j respectively.

$$D_{SL}(C_r, C_s) = \min_{\substack{a_i \in C_r \\ a_j \in C_s}} d(a_i, a_j) \qquad (1)$$

Although single linkage method is robust to small changes of individual data points, due to its clustering process, most of the time it produces ineffective clusters due to a problem known as chain or bridging problem where clusters are made by the inclusion of individual data points to one big cluster each time, resulting in a "skewed and biased" dendrogram.

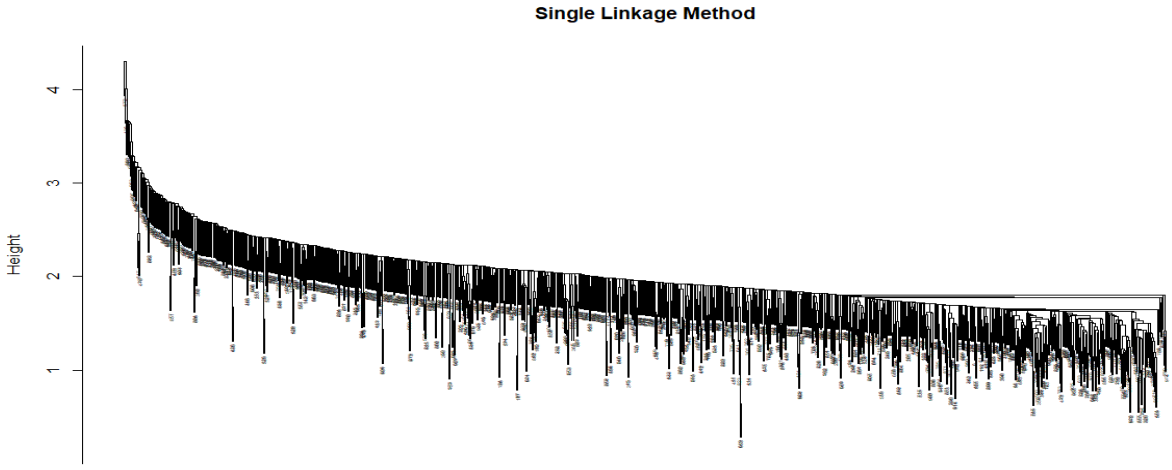

Figure 4. Dendrogram Using Single Linkage Method

From the dendrogram in figure 4, it is not easy to compare the similarities of clusters or individual data points clearly as it contains the full data of 1000 observations. A comparison of different clustering methods to increase interpretability will be provided after the examination of each method individually. One can observe the chain problem or the "skewed" dendrogram, that is typical of single linkage method and tend not to produce the most optimal result in most of the cases. Another insight

that could be derived from figure 4 is that there might be some possible outliers that stick out from the dendrogram but any insight made from the given dendrogram wouldn't be reliable as the output itself is not. Lastly, just to make sure before giving a final decision, when one checks the size of the first 10 clusters, he or she would see first cluster consisting of 990 individuals and the rest 1; however, one can also directly eliminate single linkage method just by the shape of the dendrogram as it is obvious that it won't outperform the other clustering methods, which might be listed as one of the advantages of dendrograms.

4.2 Complete Linkage Method

In the complete linkage method, the clustering is done based on the maximum distances between data points or clusters. To be more specific, the way complete linkage method merges clusters is that every time it is going to merge two clusters a new distance matrix is calculated where the elements of it consists of maximum distances of all of the possible combinations of the current clusters, data points that are not already clustered and the combination of the two. The combination that yields the smallest maximum distance between two clusters is merged so that closer or similar clusters get combined together earlier than the others. Simply put, the smallest maximum distance in each new version of the distance matrix is the height value at which the previous cluster is clustered at. Although complete linkage method and single linkage method might seem similar to each other, the dendrograms they produce differ significantly from each other as will be demonstrated below.

$$D_{CL}(C_r, C_s) = \max_{\substack{a_i \in C_r \\ a_j \in C_s}} d(a_i, a_j) \qquad (2)$$

The formal definition of how the inter-cluster distances are computed can be seen above in equation number 2. Some of the things that is good to know before applying complete linkage method are that in this method large clusters tends to grow slowly, that there might be some instability when there are small changes and lastly that it is suitable for splitting data with unclear structure.

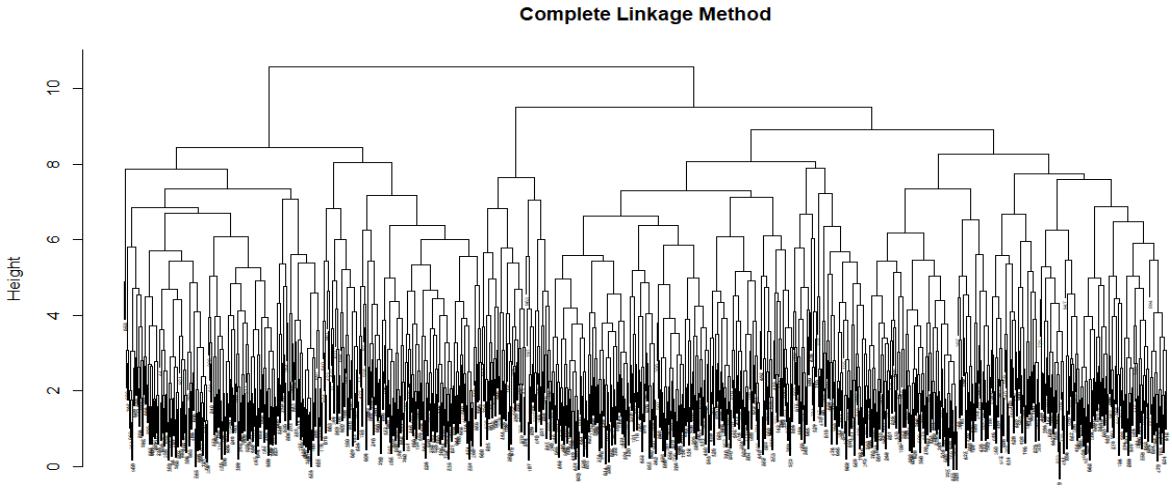

Figure 5. Dendrogram using Complete Linkage Method

Looking at the dendrogram in figure 5, a simple and immediate decision that could be taken is that the complete linkage method produces a better and somewhat more balanced clustering compared to the single linkage method. Although clusters seem to be formed relatively early in the process (low height values), as expected, big clusters are formed rather late in the process (high height values). Overall, one might keep complete linkage method for further consideration and analysis as it has some potential to produce distinct clusters.

4.3 Average Linkage Method

Average linkage method could be considered as a mix of single linkage and complete linkage methods. The reason for that is that rather than finding the minimum distance or maximum distance between clusters, whenever a clustering decision will be made, which means a new version of the distance matrix is calculated, the average of the distance combinations between the clusters' points are calculated and put in the new distance matrix. The combination that produces the minimum distance to each other are merged until there is no possible clustering option is left.

$$D_{AL}(C_r, C_s) = \frac{1}{n_r n_s} \sum_{a_i \in C_r} \sum_{a_j \in C_s} d(a_i, a_j) \quad (3)$$

$$\text{mit} \quad n_i = |\, C_i \,|$$

Looking at equation number 3, one can see the criteria based on which the new distance matrix is calculated after each iteration where, in addition to the formulas before, n_r and n_s represent the number of observations in each cluster. Since average linkage method neither uses minimum distances between observations nor the maximum but the average, it also combines the advantages and disadvantages of the first two linkage methods. The degree to which is likely to depend on the structure of the given data set.

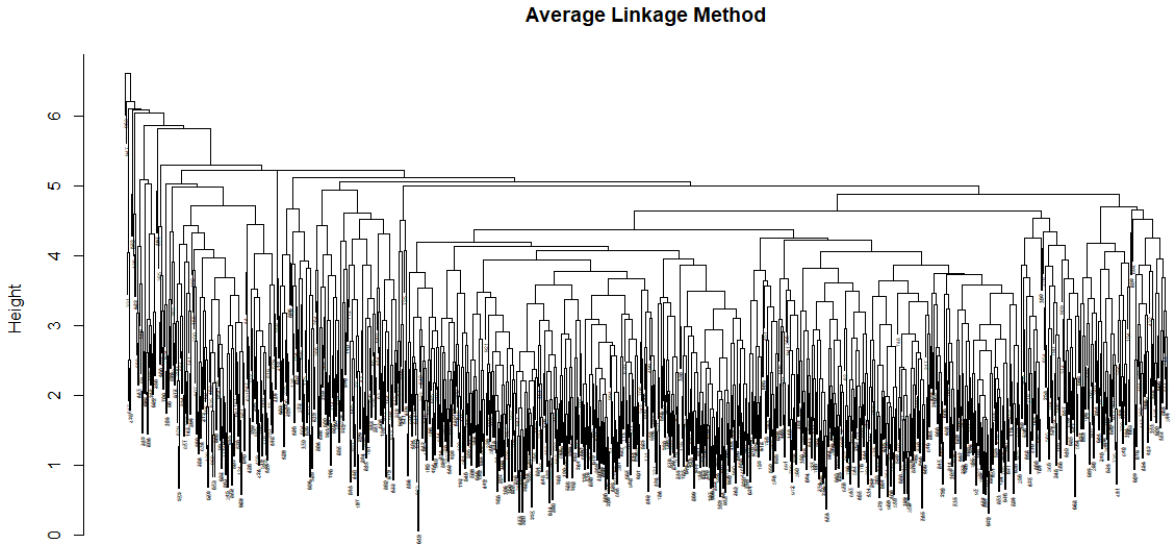

Figure 6. Dendrogram using Average Linkage Method

As expected, in figure 6, we see a combination of the typical results of single linkage and complete linkage methods, meaning that although not as strong as the figure 4, where the dendrogram was

produced using the single linkage method, we still see a slight chain-effect to the left and big clusters are merged relatively late in the process, which is a property of complete linkage method. The chain-effect is not as high as figure 4 but because the clusters' size decreases to 20s after the first two clusters of 854 and 107, it is safe to eliminate this linkage method as it doesn't create meaningful and usable clusters.

4.4 Centroid Method

In the centroid method, the clustering is based on the differences of the means of the clusters. To be more specific, in each clustering decision the elements of the distance matrix are calculated as the difference of the clusters' means, and in each iteration the smallest ones are merged until there is no more clusters to be merged. By the nature of the clustering criteria, centroid method shows some similarities to the average linkage method.

$$D_Z(C_r, Cs) = \| \bar{x}_r - \bar{x}_s \|^2$$
$$\text{where } \bar{x}_i = \frac{1}{n_i} \sum_{j \in C_i} x_j \qquad (4)$$

Taking a look in the 4th equation above, one can see the mathematical representation of how the iterative distance matrices during the clustering process are calculated. There are couple of things that one should be aware of before applying this method of hierarchical clustering. First of all, the use of the centroid method is only suitable when the data is metric. Secondly, because of the way that iterative distance matrix is calculated, the most correct application of the centroid method is only possible with the squared Euclidean distances. One possible advantage of this method could be its robustness to outliers relative to the other hierarchical clustering methods; however, when using the centroid method one should also be careful about the problem of reversals, where a data point or a cluster joins a cluster at a distance that is smaller than the previous merge of two clusters, which is against the working principle of hierarchical clustering.

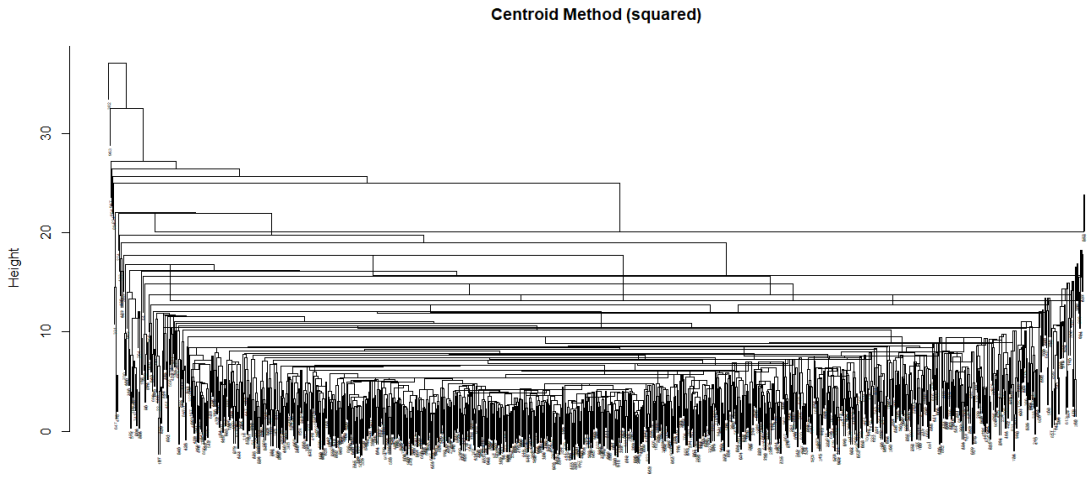

Figure 7. Dendrogram Using the Centroid Method

In figure 7, one can see the dendrogram as an output of the centroid method that used the squared distance matrix as an input, which is far away from providing useful and distinct clusters. The reason

for that is its complicated structure and low interpretability. It also suffers from the chain effect, which is something not desired in a dendrogram. Likewise the dendrogram produced by the single linkage method, one can decide on not using the centroid method by solely looking at the dendrogram as it seems problematic and obvious that it won't provide well-separated clusters.

4.5 Ward's Method

Ward's method belongs to the variance based agglomerative hierarchical clustering group among the other hierarchical clustering methods. It is an ANOVA based method and it uses the residual sum of the squares for clustering, meaning that in each clustering iteration it calculates the distance between the clusters and data points and the cluster means. Whichever combination produces the smallest distance and yields the smallest variance increase when included to a cluster, is merged until one big cluster is reached or singletons in the case of divisive hierarchical clustering.

$$D_W(C_r, C_s) = \frac{n_r n_s}{n_r + n_s} ||\bar{x}_r - \bar{x}_s||^2 \qquad (5)$$

Equation number 5 shows the objective function that is used to merge clusters using the ward's method. An advantage of the ward's method is that it is good in finding clusters when there is noise in the data set. However, although it tends to produce better results than most of the hierarchical clustering methods it has some possible disadvantages too. One of them, which is also quite important, is the fact that it does not return an optimal partition. Another is that, sometimes it might be biased towards globular clusters. Lastly, likewise the centroid method, ward method can only be applied to metric data and the possibility of reversions still exist, which is something not desired.

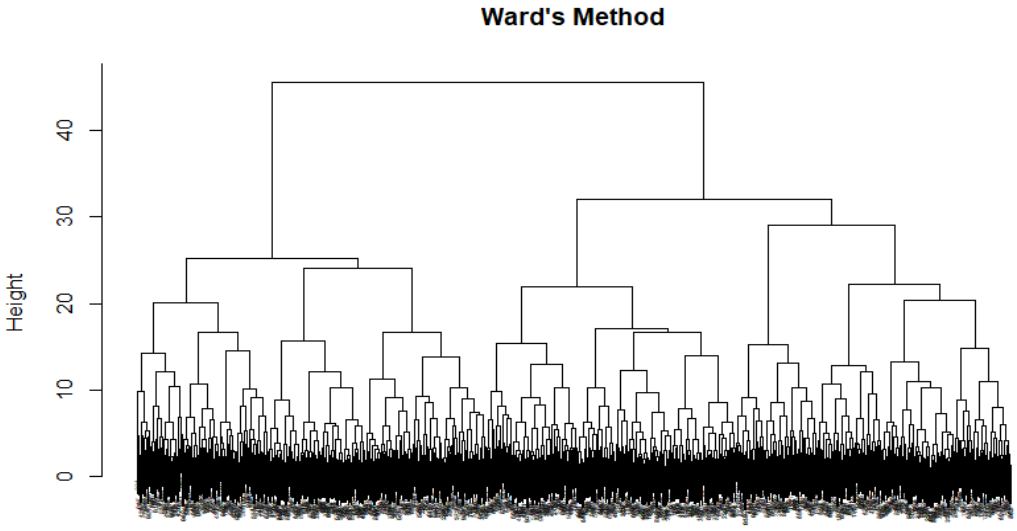

Figure 8. Dendrogram Using the Ward's method

The dendrogram in figure 8, shows the output of the Ward method. The dendrogram doesn't show any chain effect, which is a necessary step for the clustering to be usable in most cases. Also, the clusters are formed relatively early in the process (low height values) and the distribution of the clusters seems balanced or "unbiased". From a first look at the dendrogram, Ward's method is more advantageous over other methods but similar to the decision of number of clusters, the decision of

which clustering method to be chosen cannot be made only relying on the dendrogram, which will be discussed in the next and the last section.

5. Comparision of the methods and the dendrograms

As mentioned at the first part of the report, in clustering there are many decisions which can affect the clustering output. Some of these decisions are which distance measure will be used (Euclidean, gower, Mahalanobis, squared distance etc.) and how the variables will be scaled. The given data set was produced randomly and the variables don't have any meaning but in real-life scenarios one can try each clustering method with a combination of different distance measures, scaling techniques, number of clusters and different variables, which in total creates quite some number of possibilities considering the number of clustering methods there exist such as k-means, model-based clustering methods, density-based clustering methods etc. if one wants to really find the "optimal" clustering for the specific case at hand. One additional approach that one could try given a real data set is to use Principal Component Analysis, to decrease the dimensionality of the data set while keeping as much information as possible that are in the variables. Also, contrary to an easy to do mistake for beginners, dendrograms neither gives an answer to the question of how many clusters will there be nor to the one that which clustering method is the best one. These are decisions that needs to be made by the researcher with the consideration of the possible decision combinations mentioned above. So, although there are some expected outputs with some methods like chain effect in the single linkage method and these can be eliminated rather early in the process of giving clustering decisions, still to come up with a good enough decision it is likely that one would need to iterate through at least some of the possible clustering decision combinations by comparing the results of each of them with each other.

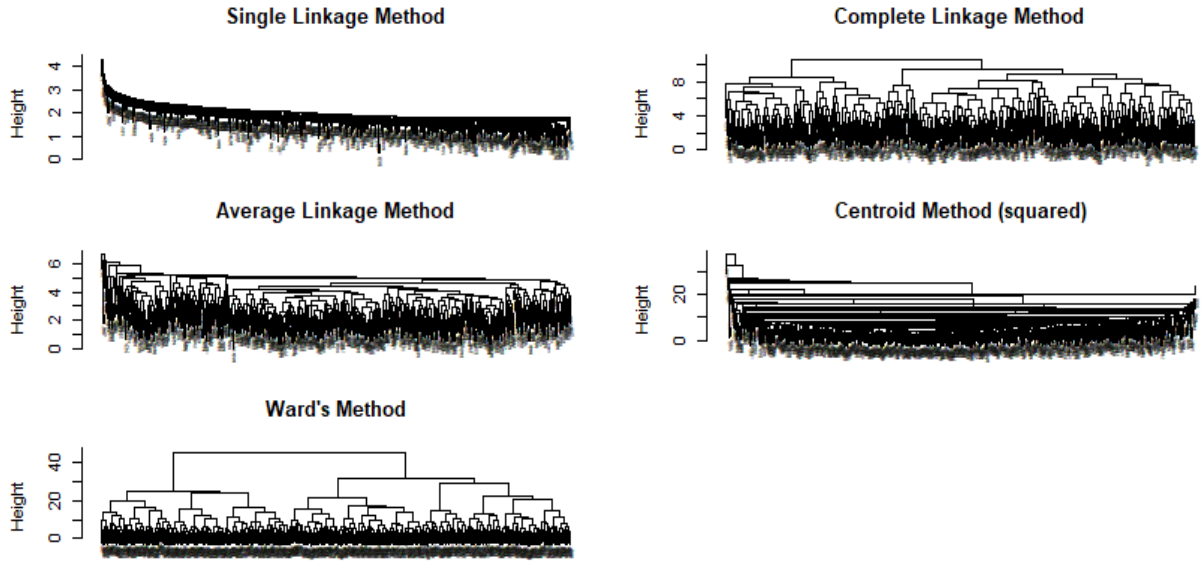

Figure 9. Comparison of the dendrograms

In this report the focus was on the calculation and the interpretation of the dendrograms. Because of that reason, iterations with different dendrograms using different distance metrics or combination of variables and other parameters were not done as this would have only pointed out the difference

between clustering decisions, as well as it would only shift the focus of the report and wouldn't contribute to dendrogram interpretations. In the case of the given data set and the dendrograms of the 5 different methods as can be seen in figure 9, which clustering method would be chosen could be decided fairly easier. As single linkage, average linkage and the centroid method produced chain effect, they were excluded from further consideration. To decide on whether Ward method or complete linkage method will be used, contextualization of the data set and further examination of the outputs produced by both of the methods needs to be done. To sum up, to make a final decision, one needs to consider the variation in the clusters, the values of the variables in them, while the decision could be made easier using a combination of other multivariate visualization techniques than dendrograms such as MDS plot of the data points and the clusters or parallel coordinate charts or fit values.

REFERENCES

[1] Chapmann, C., Feit, E. M. (2019). R for Marketing Research and Analytics (2nd ed). Springer. 10.1007/978-3-319-14436-8

[2] Dabhi D. P., Patel M. R. (2016). *Extensive Survey on Hierarchical Clustering Methods in Data Mining*. 3(11). 659-665. Retreived from: https://www.irjet.net/archives/V3/i11/IRJET-V3I11115.pdf

[3] Everit, B. S., Dunn, G. (2001) *Applied Multivariate Data Analysis* (2nd ed). Wiley. 10.1002/9781118887486.

[4] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2019). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0.

[5] NCSS LLC. (n.d.). *Clustering in NCSS*. Retrieved March 18, 2021, from https://www.ncss.com/software/ncss/clustering-in-ncss/

[6] Jarman, A. (2020). Hierarchical Cluster Analysis: *Comparison of Single linkage,Complete linkage, Average linkage and Centroid Linkage Method*. 10.13140/RG.2.2.11388.90240.

[7] Saraçli, S., Doğan, N. & Doğan, İ. *Comparison of hierarchical cluster analysis methods by cophenetic correlation*. J Inequal Appl 2013, 203 (2013). https://doi.org/10.1186/1029-242X-2013-203

[8] Shalizi, C. (2009) *Distances Between Clustering, Hierarchical* Clustering [PDF] Retrieved from 36-350 Data Mining Course, Carnegie Mellon University. https://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf

APPENDIX

The R code used in the analysis (version 4.0.3):

```
setwd("C:/Users/Administrator/Downloads")
library(cluster)
library(ggplot2)
library(factoextra)
library(dendextend)
library(idendr0)

data <- read.csv("Mert.csv")



column_names <- c("1", "2", "3", "4", "5", "6", "7", "8", "9", "10","11","12")
colnames(data) <- column_names

summary(data)

data_scaled <- scale(data,scale=TRUE,center=TRUE)
data_dist <- dist(data_scaled) # compute distance on standardized data

hc_single   <- hclust(data_dist, method="single")
plot(hc_single, main = "Single Linkage Method", cex=0.3)
hcd_single <- as.dendrogram(hc_single)
plot(cut(hcd_single, h = 3.25)$upper, main = "Upper tree of 10 clusters, Single Linkage")
#low interpretability but shows that doesn't produce a good result

dend <- hcd_single
dend <- color_branches(dend, k=10)   #shows how unproportional the clusters are
```

```
plot(dend)

hcd_single_cut <- cutree(hc_single, k = 10)   #chain problem
table(hcd_single_cut)
plot(rev(hc_single$height^2))

hc_complete   <- hclust(data_dist, method="complete")
plot(hc_complete, main = "Complete Linkage Method", cex=0.3)
sub_grp <- cutree(hc_complete, k = 10)
table(sub_grp)
plot(rev(hc_complete$height^2))
fviz_dend(
  hc_complete,
  k = 10,
  horiz = FALSE,
  rect = TRUE,
  rect_fill = TRUE,
  rect_border = "jco",
  k_colors = "jco",
  cex = 0.1, main = "Complete Linkage Method")

hc_average   <- hclust(data_dist, method="average")
plot(hc_average, main = "Average Linkage Method", cex=0.3)
sub_grp <- cutree(hc_average, k = 10)
table(sub_grp)   #chain effect
plot(rev(hc_average$height^2))

hc_centroid   <- hclust(data_dist, method="centroid")
```

```
plot(hc_centroid, main = "Centroid Method", cex=0.3)
sub_grp <- cutree(hc_centroid, k = 10)
table(sub_grp) #chain effect
plot(rev(hc_centroid$height^2))

hc_centroid_squared   <- hclust(data_dist^2, method="centroid") #this version is used in the report
plot(hc_centroid_squared, main = "Centroid Method (squared)", cex=0.3)

hc_ward   <- hclust(data_dist, method="ward.D2")
plot(hc_ward, main = "Ward's Method", cex=0.3)
sub_grp <- cutree(hc_ward, k = 10)
table(sub_grp)
plot(rev(hc_ward$height^2))

fviz_dend(
  hc_ward,
  k = 10,
  horiz = FALSE,
  rect = TRUE,
  rect_fill = TRUE,
  rect_border = "jco",
  k_colors = "jco",
  cex = 0.1, main = "Ward's Method")

hc_ward_4 <- cutree(hc_ward, k=4)

seg.summ <- function (data , groups)     #could be useful if the variables had meanings
{aggregate (data , list(groups), function (x) mean(as.numeric (x)))}
```

```
tmp <- seg.summ(data, hc_ward_4)        #one could also check the variances in each cluster's variables

tmp


dend_plot <- fviz_dend(hc_ward)               # create full dendogram

dend_data <- attr(dend_plot, "dendrogram") # extract plot info

dend_data

dend_cuts <- cut(dend_data, h = 18)      # cut the dendogram

dend_cuts  #shows where at which heigt the clusters were merged

hcd_ward <- as.dendrogram((hc_ward))

plot(cut(hcd_ward, h = 18)$upper, main = "Upper tree of 10 clusters")

#High interpretability. The interpretation on the similarity of the "Branches" is done

#in figure 1,2 and 3.
```

Declaration

I hereby confirm that I have authored this document independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

I have attached the code used to produce the analysis in the appendix. I confirm that I have written and executed the analysis, and that the code is complete and executable.

Turkey, 19/03/2021

PLACE, DATE

SIGNATURE

Mert Erdinc

.....

Your NAME