Hierarchical Clustering and the Determination of Maturity in Trees

Ganiyu A. Dawodu, Ph.D.[1]; Isqeel A. Ogunsola, B.Sc.[2]; and Osebekwin. E. Asiribo, Ph.D.[3]

Department of Statistics, Federal University of Agriculture, Abeokuta, Nigeria.

E-mail: agad1963@yahoo.com[1], dawoduga@funaab.edu.ng[1], ogunsolaia@funaab.edu.ng[2], asiribooe@funaab.edu.ng[3]

ABSTRACT

The determination of the maturing of a tree lies on the volume of wood contained in it. The said volume has been found to depend on the girth and height of the tree and cannot be determined by merely looking at the tree. In this work, hierarchical clustering was used in the process of identifying all the clusters and their respective cluster sizes to facilitate the graduation of the maturity of two species of trees from the lowest cluster to the highest.

(Keywords: Euclidean, forest management, girth, hierarchical clustering, tree harvesting)

INTRODUCTION

Natural resources are in limited supply throughout the countries of the world; they therefore ought to be utilized judiciously. An aspect of the natural resources of a country is associated with the preponderance of its forests. Deforestation is a scourge that affects people in at least two ways, first of which is the removal of a major source of replenishment of the atmospheric oxygen; the second, has to do with the topic of the present research, and is the untimely felling of immature trees. This often results in the felling of many trees instead of one or two if these trees were allowed to mature before harvesting them.

Clustering generally constitutes a set of techniques for automatic classification of samples into a number of groups using a measure of association. The samples in one group are as homogeneous as possible whilst the samples in different groups are supposed to be as heterogeneous as possible. Clustering is a practical approach and it is applicable in many areas, including artificial intelligence, medical research, and geosciences, which is the main concern in this work (Kantardzic, 2003; Kaufman and Rousseeuw, 2005).

The major requirement for clustering is a set of samples and, at least, a measure of similarity (or dissimilarity) between the samples. The results from clustering will often be a dendrogram which will contain a number of groups (clusters) that form a partition, or a structure of partitions, of the data set. One additional result of clustering is a generalized description of every cluster, and this is especially important for a detailed analysis of the data set's characteristics.

Samples for clustering are represented as row vectors of measurements, that is, as points in a multidimensional space. In the results, samples within a dendrogram cluster are more similar to each other than those belonging to different clusters. Clustering methodology is particularly appropriate for the exploration of interrelationships among samples to make a preliminary assessment of the sample structure.

Humans perform competitively with automatic-clustering procedures in one, two, or three dimensions, but most real-life problems involving clustering are in higher dimensions (i.e., greater than three) thus making it difficult for a human being to classify. It is very difficult for humans to intuitively interpret data embedded in a high-dimensional space. Altogether there are three types of clustering algorithms, namely; Hierarchical (Agglomerative), Partitional, and Incremental. The trio requires data of samples to work. Each sample (or row entry) contains measures of similarities, s_i (or dissimilarities, d_i) that run over a constant number of columns.

With respect to the data that are used for the present work, the species of trees are, *Triplochiton scleroxylon* (Obeche) and *Terminilia ivorensis* (Black Afara), each having

its stand at Onigambari forest reserve, Ibadan, Oyo State, Nigeria.

Each row entry of the data, $X = \{S_i\}$ *or* $X = \{d_i\}$, $i = 1, 2, \ldots, n$, running over $n+1$ columns, keeps the first row mainly for identification. After clustering, a dendrogram, D, creates all the clusters in the form of partitions, H_j, $j = 1, 2, \ldots, m$, such that:

$$D = \begin{cases} \bigcup_{j=i}^{m} H_j, & m \leq n \\ H_k \cap H_i = \emptyset, & \forall\, k \neq t \end{cases} \quad \textbf{(1)}$$

MATERIALS AND METHODS

By assuming that tree logs are roughly cylindrical, in shape, then volume of wood in each log, V, in relation to girth (i.e., circumference of the tree trunk at an average human height), g, and the height of tree, h, is given by (Crawley, 2007):

$$V = \frac{1}{4\pi} g^2 h. \quad \textbf{(2)}$$

The measurements obtainable at Onigambari tree stands are those of girth and height, in meters, per tree. Here tree maturity is with respect to the volume (or quantity) of wood. Hence, by using a package in Microsoft suites (i.e. Microsoft Excel), another column of data entries, that estimates the volume (in m^3) of wood in each tree species, is created. An extract of the updated data (Table 1), on Obeche, looks like this;

Table 1: An Extract of the Updated Data on Obeche.

SN	Tree (ID)	Height (m)	Girth (m)	Volume (m^3)
1	t1	41.7	1.65	9.03
2	t2	76.0	2.07	25.93
3	t3	59.5	2.04	19.69
4	t4	62.0	2.28	25.67
5	t5	59.5	1.97	18.43
6	t6	61.7	1.89	17.51

Among the R software packages is **cluster** (Maechler et al., 2016), there exists a function known as **AGNES**, it is of the agglomerative hierarchical type, it is capable of synthesizing a sequence of clusters, the group average method is the default because it is favored whenever, the arguments possesses the properties of monotonicity, consistency, and robustness (Kaufman and Rousseeuw, 2005; Struyf et al., 1996).

Although four other well-known methods are available in AGNES, namely; single linkage, complete linkage, Ward's method, and weighted average linkage, the present work is centered around a simple invocation of AGNES. These five methods can be described in a unified way (Lance and Williams, 1966) and are already implemented in many software packages. In comparison with other implementations (Fraley and Raftery, 2002; Fraley et al., 2009; Sarkar, 2008; Sarkar, 2009), two additional features are put into the R package, cluster, which is made to; yields the agglomerative coefficient which measures the amount of clustering structure found, and apart from the usual tree dendrogram, it also provides the banner, which is a graphical display.

The Agglomerative Coefficient (AC) (Rousseeuw, 1986) is a quality index for an agglomerative cluster of the data, it suffices it to say that the closer to 1.0 (100%) an AC is, the better and reliable, the pertinent analysis is. In R, the invocation of AGNES is:

AGNES (x, diss = inherits (x, "dist"), metric = "euclidean", stand = FALSE, method = "average", par.method, keep.diss = n < 100, keep.data = !diss, trace.lev = 0) **(3)**

With respect to the "metric" clause in the invocation of AGNES, (3), the dissimilarity coefficients between objects may be obtained from the computation of distances, such as respectively, Euclidean (4) or Manhattan (5) distances:

$$E(i,j) = \sqrt{(d_{i1} - d_{j1})^2 + (d_{i2} - d_{j2})^2 + \cdots + (d_{in} - d_{jn})^2} \quad \textbf{(4)}$$

or

$$M(i, j) = |d_{i1} - d_{j1}| + |d_{i2} - d_{j2}| + \cdots + |d_{in} - d_{jn}| \quad \textbf{(5)}$$

Between the objects with measurements, d_{i1}, d_{i2}, …, d_{in} and d_{j1}, d_{j2}, …, d_{jn}. Sometimes researchers use similarities, $\{S_i\}$, $i = 1,2,\ldots, n$, taking dichotomous values, $\{0,1\}^n$, but they are strongly advised to transform them into dissimilarities. $E(i,j)$ is more popular than $M(i,j)$ because it is similar to the distance between two Euclidean, n R^n vectors:

$$\bar{x}_i = \{d_{ik}\}, \bar{x}_j = \{d_{jk}\}, \quad k = 1,2, \ldots., n \quad \textbf{(6)}$$

The *M(i,j)* was first used in clustering context by Carmicheal and Sneath (1969), is the preferred metric if locations are perfectly North-South or East-West to each other.

With respect to the data (Table 1), it is possible to compute a dissimilarity matrix from these data (entries from specific trees). For instance by calculating Euclidean distances between objects (i.e. trees) we obtain the symmetric square-matrix, $M_x\ (d_{ij})$ *, i, j, = 1,2, …, n*, whose diagonal entries are all equal to zero (since they each measure the distance of a tree to itself, that is, *d(i,i) = 0, (i = 1,2,…, n).* Consequently, the dissimilarity between tree, i and tree, j will be located at the intersection of the i^{th} row and the j^{th} column of $M_x\ (d_{ij})$ *, i, j, = 1,2, …, n*.

The creation of clusters will now begin by ranking the distinct distance metric inside the dissimilarity matrix from the least to the highest to obtain a set H_K, k =1,2,…,m. This will enable us to create as much as m clusters if we merely leave all trees that are equal-distant from the first tree in a cluster (while leaving the first tree in the first cluster only) and thus dendrogram (i.e., equation (1)), D is created. Sometimes m may still be too large (this is particularly true when almost all of the clusters have just one element), we may decide to make a contraction of it through the following approach.

Let; C_1, C_2 be any two consecutive clusters, $|C_1|$, $|C_2|$ be the number of trees in them respectively and $\bar{d}(i,j)$, $i \in C_1$, $j \in C_2$, be the average distance metric between the trees in clusters C_1 and C_2. Then we can create yet another dissimilarity matrix, $M_D\ (C_{ij})\ i \in C_1,\ j \in C_2$, such that:

$$\bar{d}(i,j) = \frac{1}{|C_1|,|C_2|} \sum_{\substack{i \in C1 \\ j \in C2}}^{\cdot} d(i,j) \qquad \textbf{(7)}$$

We can now use the newly created $M_D\ (C_{ij})\ i \in C_1,\ j \in C_2$, to create a new set of clusters with each cluster having more trees than in the enlargement, *D* , albeit with the number clusters reduced. The decision to contract further may come in if the number of clusters is still large.

RESULTS

The implemented package in R (Maechler et al., 2016), known as **cluste**r that was written from its Fortran equivalent (Struyf et al., 1996) contains all the subroutines for concepts that have been explained in Materials and Methods section. To illustrate how the package works, the data from Onigambari tree stands of Obeche and Black Afara will be used. Concerning the data on Obeche, the following, short invocation, will efficiently analyze it (i.e., by assuming that the datafile, "Bayes1.csv" is left in a folder named "Dawodu" on the Desktop).

```
library(cluster)
Triplochiton=
read.table("C:\\Users\\FUNAAB\\Desktop\\DAWO
DU\\Bayes1.csv",
header=TRUE,sep=",")
Obeche <- agnes(Triplochiton, metric =
"manhattan", stand = TRUE)
plot(obeche)
```

To obtain the results:
Call: agnes(x = Triplochiton, metric = "manhattan", stand = TRUE)
Agglomerative coefficient: 0.9095036

Order of Objects: [1] 1 14 104 120 105 106 7 78 56 18 23 38 32 40 13 119 16 21 25 28 115 48 72 84 85 89 90 86 36 51 74 82 2 41 49 55 62 67 57 46

[41] 4 45 53 31 80 91 60 63 3 22 24 19 15 110 113 107 100 101 111 121 102 114 10 11 109 118 5 58 47 6 44 69 77 64 68 75 73 61 81 66

[81] 99 8 79 59 94 96 9 92 87 93 50 52 65 71 12 116 27 26 29 39 33 42 112 117 35 37 122 30 34 108 43 54 83 88 95 97 98 70 76 17

[121] 103 20

Height (Summary):

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.09	0.75	1.18	1.74	2.19	11.19

Available Components:

[1] "order" "height" "ac" "merge" "diss" "call" "method" "data"

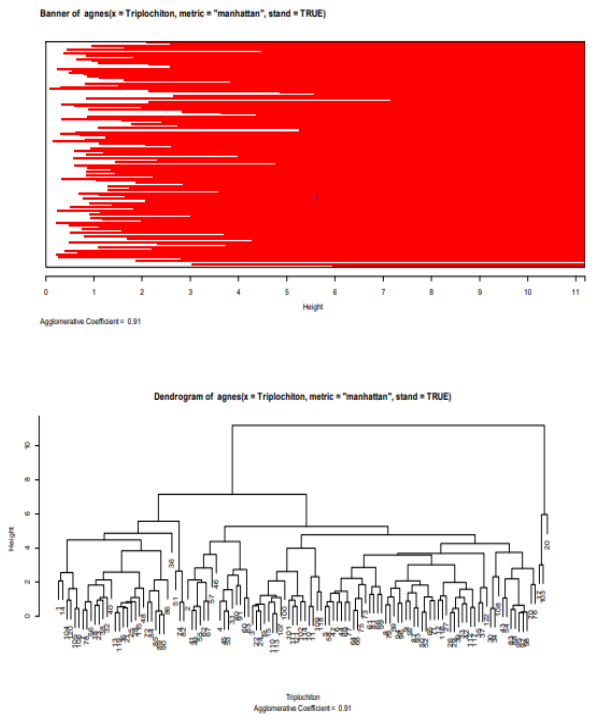

Figure 1: The "Order of Objects", "Height (Summary)", Banner and Dendrogram Obtained on the Data on Obeche.

By repeating the process over the data on Black Afara whose extract is shown in Table 2.

Table 2: An Extract of the Updated Data on Black Afara Tree Stand.

SN	Tree (ID)	Height (m)	Girth (m)	Volume (m^3)
1	T1	13	1.05	1.15
2	T2	16	1.05	1.41
3	T3	16	0.99	1.25
4	T4	15	0.65	0.51
5	T5	15	0.66	0.52
6	T6	18	0.87	1.09

We obtain; Call: agnes(x = Terminilia, metric = "manhattan", stand = TRUE)
Agglomerative coefficient: 0.8757367

Order of Objects:
[1] 1 2 25 43 3 34 41 30 21 33 24 6 59 65 62 64 66 74 7 50 57 63 4 5 52 35 36 39 10 14 28 20 23 27 32 29 40 44 42 51 54 58 60 56 53 17 18 19 38 45 46 48 49 68

[55] 70 72 69 71 37 9 75 55 61 8 73 47 67 76 77 80 78 79 31 11 12 16 13 15 22 26

Height (Summary):

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	0.89	1.34	1.88	2.30	8.94

Available Components:

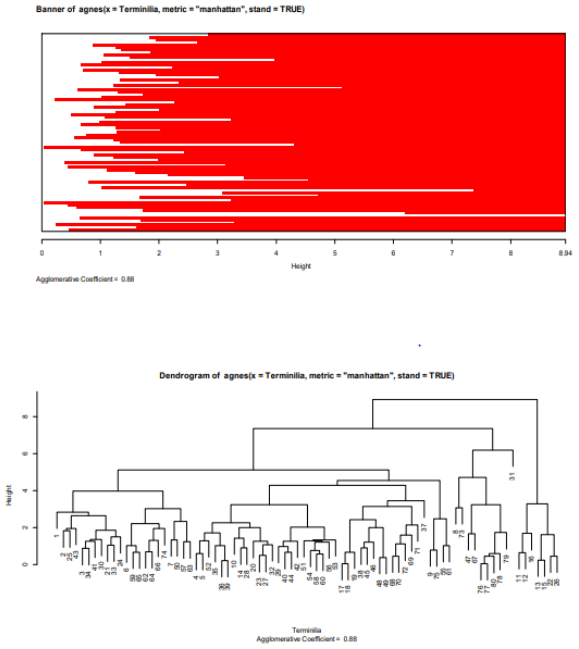

Figure 2: The "Order of Objects", "Height (Summary)", Banner and Dendrogram Obtained on the Data on Black Afara.

DISCUSSION

The short invocation of AGNES is reusable whenever the datafile is in the excel "csv" format and left in a folder on the desktop or any easily accessible folder (e.g., documents), the analyst however needs to be sure that the "address" in the second line of the invocation is properly directed to the datafile.

A part of the botanical and local names are used in the invocation to enable analysts that want to reuse it to know where to change things without affecting the functionality of the invocation. Since a tree's maturity is with respect to the volume (or quantity) of wood and not with respect to its age, the column containing the volume of wood becomes quite useful in the determination (i.e., determination of the beginning and end) of the clusters.

With respect to the second tree (Black Afara), the; data size is just 80, agglomerate coefficient is 0.8757367 (or approximately 88%). For the sake of brevity, its result will now be used to identify all its clusters (and consequently its most matured cluster). There will be a sudden significant difference between the volume of wood quantity of the last tree in the last cluster and the first tree in the new cluster as the generation of clusters continues. The first cluster, having tree identifications, 1, 2, 25, 43, 3, 34, 41, 30, 21, 33, 24, 6, 59, 65, 62, 64, 66, and 74, is contained in Table 3 below:

Table 3: The First Cluster on Black Afara has Cluster Size, 1.

Tree Id	Height	Girth	Volume
T1	13	1.05	1.15
T2	16	1.05	1.41
T25	15	1.17	1.64
T43	15	1.08	1.39
T3	16	0.99	1.25
T34	16	0.93	1.11
T41	15	0.97	1.12
T30	14	1.01	1.14
T21	15	0.92	1.01
T33	15	0.88	0.92
T24	13	0.93	0.89
T6	18	0.87	1.1
T59	16.5	0.91	1.09
T65	16.5	0.88	1.02
T62	15.5	0.99	1.2
T64	14.5	1.01	1.17
T66	14	1.09	1.32
T74	14.5	0.91	0.96

Similarly, the second cluster is contained in Table 4 and it has cluster size, 17 (because the difference in volume of the last tree in the last cluster and the first tree in the new cluster, i.e. 0.96 – 0.65 = 0.31 is rather large):

Table 4: The Second Cluster on Black Afara has Cluster Size, 17.

Tree Id	Height	Girth	Volume
T7	17	0.69	0.65
T50	17	0.72	0.71
T57	13.5	0.83	0.74
T63	15	0.79	0.74
T4	15	0.65	0.51
T5	15	0.66	0.52
T52	14.5	0.74	0.63
T35	15	0.57	0.39
T36	13.5	0.61	0.4
T39	13.5	0.62	0.42
T10	14	0.55	0.34
T14	14	0.65	0.47
T28	14	0.67	0.5
T20	14	0.83	0.77
T23	13.5	0.74	0.59
T27	13.5	0.77	0.64
T32	14.5	0.77	0.68

Continuing in this fashion, Table 5 contains the third cluster:

Table 5: The Third Cluster on Black Afara has Cluster Size, 9.

Tree Id	Height	Girth	Volume
T29	11.5	0.52	0.24
T40	11.5	0.51	0.24
T44	12	0.54	0.28
T42	11	0.43	0.16
T51	14	0.52	0.3
T54	12.5	0.52	0.27
T58	12	0.59	0.33
T60	12.5	0.56	0.31
T56	13	0.44	0.2

Table 6 contains the fourth cluster; Table 7 contains the data for the fifth cluster; Table 8 contains the data for the sixth cluster; and likewise, Table 9 contains the data for the seventh cluster,

Table 6: The Fourth Cluster on Black Afara has Cluster Size, 15.

Tree Id	Height	Girth	Volume
T53	13	0.64	0.43
T17	11	0.9	0.7
T18	11	0.9	0.7
T19	12	0.87	0.72
T38	13	0.88	0.8
T45	13	0.82	0.7
T46	12	0.93	0.83
T48	10	0.85	0.57
T49	10.5	0.85	0.61
T68	11.5	0.86	0.68
T70	11	0.88	0.68
T72	10.5	0.96	0.77
T69	12	1	0.95
T71	8.5	0.88	0.52
T37	10	1.12	0.99

Table 7: The Fifth Cluster on Black Afara has Cluster Size, 4.

Tree Id	Height	Girth	Volume
T9	8	0.55	0.19
T75	8.5	0.58	0.23
T55	10	0.69	0.38
T61	9	0.67	0.32

Table 8: The Sixth Cluster on Black Afara has Cluster Size, 10.

Tree Id	Height	Girth	Volume
T8	17	1.32	2.36
T73	13	1.46	2.21
T47	10	1.39	1.53
T67	10.5	1.34	1.49
T76	12	1.16	1.28
T77	12	1.16	1.28
T78	12.5	1.18	1.38
T79	12.5	1.3	1.67
T80	12	1.18	1.33
T31	10.5	1.62	2.19

Table 9: The Seventh Cluster on Black Afara has Cluster Zize, 7.

Tree Id	Height	Girth	Volume
T11	8	0.4	0.1
T12	8	0.5	0.16
T16	7.5	0.66	0.26
T13	4.5	0.39	0.06
T15	4.5	0.36	0.05
T22	4	0.18	0.01
T26	4	0.24	0.02

CONCLUSION

In conclusion, the most matured cluster (with the most matured tree having volume of wood, 2.36 m^3) is the sixth whilst the least matured cluster (with the least matured tree having volume of wood, (0.01m^3) is the seventh. But, of course, as suggested by the AC value of 88%, this is still not perfect and hence under thorough inspection, it can be easily seen that the third and fifth clusters ought to have been merged.

From this discussion, it is apparent that the cluster could be graduated from that of the most matured cluster (consequently, most matured tree) down to the least matured cluster (consequently, the least matured tree). In order to exhibit the clusters, the result of the analysis was used to fetch the full data of the particular tree whose identity is next in the result, hence in order to further simplify the work the result of the automation of the process of exhibiting the clusters that has been kept diagrammatically in the dendrogram should always be utilized, the only problem that may disturb this is that when the data size is very large (say above 100) the dendrogram may become very clumsy even when the printing is set to portrait. In that case the analyses may be done in batches.

REFERENCES

1. Carmicheal, J.W. and P.H.A. Sneath. 1969. "Taxometric Maps". *Systematic Biology*. 18(4): 402-415.

2. Crawley, M.J. 2007. *The R Book*. John Wiley and Sons Ltd. West Sussex, UK.

3. Fraley, C. and A.E. Raftery. 2002. "Model-Based Clustering, Discriminant Analysis, and Density Estimation". *Journal of the American Statistical Association*. 97:611-631.

4. Fraley, C., A.E. Raftery, and R. Wehrens. 2009. "mclust: Model-based Cluster Analysis". URL http://www.stat.washington.edu/mclust, R package version 3.1-10.3.

5. Kantardzic, M.J.B. 2003. *Data Mining Concepts, Models, Methods and Algorithms*. IEEE Press Wiley-Interscience. A John Wiley and Sons, Inc.: New York, NY.

6. Kaufman, L. and P.J. Rousseeuw. 2005. *Finding Groups in Data; An Introduction to Cluster Analysis*. John Wiley and Sons, Inc.: Hoboken, NJ.

7. Lance, G.N. and W.T. Williams. 1966. "A General Theory of Classificatory Sorting Strategies: Hierarchical Systems". *Computer J.* 11:195.

8. Maechler, M., P. Rousseeuw, A. Struyf, and M. Hubert. 2016. "Package Cluster. Version 2.0.5. Maintainer". Martin Maechler. maechler@stat.math.ethz.ch.

9. Sarkar, D. 2008. *Lattice: Multivariate Data Visualization with R*. Springer-Verlag: New York, NY.

10. Sarkar, D. 2009. "Lattice: Lattice Graphics". URL http://CRAN.R-project.org/package=lattice, R package version 0.17-22.

11. Struyf, A., M. Hubert, and P.J. Rousseeuw. 1996. "Clustering in an Object-Oriented Environment". *Journal of Statistical Software*. 1. http://www.jstatsoft.org/v01/i04.

12. Rousseeuw, P.J. 1986. "A Visual Display for Hierarchical Classification". *Data Analysis and Informatics*. 4. North-Holland, Amsterdam. Pgs. 743-748.

ABOUT THE AUTHORS

Ganiyu A. Dawodu, is a Senior Lecturer in the Department of Statistics, Federal University of Agriculture, Abeokuta, Nigeria. He holds a Ph.D. in Statistics. His research interests are in probability distributions and statistical modelling.

Isqeel A. Ogunsola, is a Graduate Assistant in the Department of Statistics, Federal University of Agriculture, Abeokuta, Nigeria. He is currently a Master student in the Department. His research interest are in applied statistics, biometrics, and distribution theory.

Osebekwin E. Asiribo, is a Professor in the Department of Statistics, Federal University of Agriculture, Abeokuta, Nigeria. He holds a Ph.D. in Experimental Design. His research interests are in statistical modelling and experimental design.

SUGGESTED CITATION

Dawodu, G.A., I.A. Ogunsola, and O.E. Asiribo. 2020. "Hierarchical Clustering and the Determination of Maturity in Trees". *Pacific Journal of Science and Technology*. 21(1):121-127.