Journal of Modern Mathematics and Statistics 3 (3): 51-55, 2009
ISSN: 1994-5388

Cluster Analysis of Improved Cassava Varieties Cultivated at Onne, Nigeria by the International Institute of Tropical Agriculture

Nwabueze Joy Chioma
Department of Statistics, Abia State University, Uturu, Abia State, Nigeria

Abstract: Secondary data on proximate composition of fufu flour taken from 43 Cassava Mosaic Disease (CMD) resistant varieties were used for this research. Agglomerative hierarchical cluster analysis was performed on the squared Euclidean distance matrix. The distance coefficients generated between the 43 CMD resistant varieties ranged from 0.000-89.120. Six distinct groups were identified at 0.97 coefficients. A dendrogram of the data indicated that cases with low distances are close together with a line linking them. It was observed that the line was a short distance from the left of the dendrogram indicating that they were agglomerated in a cluster at a low distance coefficient. This indicated likeness. The implication of this to the farmer and indeed to the nutritionist is that a variety can be selected from each of the six cluster groups for cultivation with the objective of achieving the same nutritional differences in terms of proximate composition without having to examine all the 43 varieties.

Key words: Cluster, agglomerative hierarchical cluster, squared euclidean distance, cassava fufu, CMD-resistant varieties, clustering algorithms

INTRODUCTION

Although, cassava is an established commercial crop in many countries with 100s of varieties in existence, little is generally known of the nomenclature and identification of varieties. Various varieties are usually differentiated from one another by their morphological characteristics, such as colour of stems, petioles leaves etc. Cassava varieties are usually grouped into two main categories *Manihot palmata* and *Manihot aipi*, or bitter and sweet cassava. This grouping is a matter of economic convenience, as it is difficult to distinguish the two groups by botanical characteristics.

The Collaborative Study on Cassava in Africa (COSCA) showed that between 1961 and 1999, total cassava production in Africa nearly tripled from 33 million tons year^{-1} from 1995-1999 in contrast to the more moderate increase in Asia and Latin America (Nweke *et al*., 2002). Nigeria is currently one of the largest producers of cassava in the world with an annual output of over 45 million roots (FAO, 2002). As a result of increase in cassava production in Nigeria, many improved cassava varieties have been developed.

The cassava varieties were bred for high yield, pest disease resistant, good product quality and early maturity. The cassava varieties used in this research were developed for pest and disease resistance against the attack of common cassava disease known as Cassava Mosaic Disease (CMD), a viral disease transmitted by a white fly vector (IITA, 2005). These varieties were cultivated at Onne, by the International Institute of Tropical Agriculture (IITA) and at Umudike, Nigeria, by the National Root Crops Research Institute Nigeria. Etudaiye *et al*. (2009) processed 44 of these varieties from the two locations into fufu flours and reported a set of data on their proximate composition.

The data on the proximate composition of the varieties cultivated at Onne, Nigeria were subjected to cluster analysis-a multivariate technique for detecting natural grouping with the basic objective of data reduction. In cluster analysis, a large set of variables are reduced to a more meaningful smaller set (Crawford and Lomas, 1980) with similar objects being put in the same group. Friedman and Rubin (1967) had adopted the view that represents mixtures of multivariate normal population as a routine and basis for the design and clustering algorithms while Cunningham and Ogilvie (1972) adopted the concept of ultrametric space as a basis for the formation of cluster structure. It is expected that the algorithm will implore these spaces to recover cluster structure. Furthermore, Everitt (1993) applied the strong usual or spatial appeal to certain bivariate normal population mixtures to obtain several two dimensional plots. This study among other objectives seeks to identify the most important nutritional composition of fufu flours processed from the 43 cassava varieties and to

reduce the 43 cassava varieties into groups so that the varieties with similar nutritional composition will be in the same group.

The thrust of the study, therefore is to generate information that will arm the farmer and indeed to the nutritionist select a variety from each of the identified cluster groups for cultivation and utilization, rather than experiment on as many as the 43 varieties in other to achieve the same nutritional goal in terms of proximate composition of their fufu flours. This will be cost effective and time saving.

MATERIALS AND METHODS

Data used for the study: Secondary data on proximate composition of fufu flours processed from 43 different cassava mosaic diseases-resistant varieties were used for this research.

The amount in percentage of the proximate composition of fufu flours made from each of these cassava varieties was measured and recorded by Etudaiye *et al.* (2009). These measurements included moisture, protein, ash, fat, fiber, carbohydrate and dry matter (Table 1).

Theoretical frame work: Given the items and measurements, the basic requirements of cluster analysis include a quantitative scale for association between objects, which represent a measure of distance, which could be based on similarity or proximity, a clustering criterion and on applicable algorithm. Agglomerative hierarchical techniques, which normally produce dendrogram were used in this study.

This method starts with the calculation of the distance of each individual to all other individuals. Groups are then formed by a process of agglomeration, which is a process where all objects are placed alone in group of one. Close groups are then gradually merged until finally all individuals are in a single group. The data usually consist of the values of p variables X_1, X_2, ..., X_p for n objects. These values are then used to produce an array of distances between the varieties given as:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(X_{ik} - X_{jk}\right)^2}$$

Where:

d_{ij} = The distance between the ith and jth varieties
X_{ik} = The value of the variable X_k for variety I
X_{jk} = The value of the same variable for variety j. These distances were then used for the grouping

Table 1: Proximate composition of fufu flours processed from CMD resistant varieties*,**

Cassava	Mc	Protein	Ash	Fat	Fiber	CHO	Dm
97/4769	9.97	1.90	0.15	0.13	0.07	87.831	90.03
99/6012	8.70	1.40	0.15	0.52	0.19	88.94	91.30
94/0561	9.00	0.35	0.25	1.13	0.03	90.24	91.00
97/0162	8.64	1.70	0.25	0.50	0.12	89.89	91.36
94/0026	9.81	0.81	0.10	0.49	0.10	88.83	90.19
96/1642	9.93	0.35	0.05	0.28	0.08	89.28	90.07
98/0510	9.88	0.70	0.20	0.40	0.02	88.08	90.12
98/0505	8.62	1.05	0.30	0.42	0.16	89.35	91.38
99/3037	8.16	1.05	0.25	0.12	0.01	90.37	89.81
98/2101	7.55	1.40	0.45	0.30	0.14	88.24	92.45
97/4763	8.211	1.05	0.45	0.05	0.02	89.62	81.79
97/2205	8.89	0.75	0.35	0.11	0.20	88.64	91.13
98/1565	8.64	1.85	0.15	0.22	0.01	88.98	91.36
92/0325	9.75	1.75	1.12	0.42	0.08	87.61	90.32
92/00168	8.31	0.35	1.50	0.01	0.20	90.21	91.69
TME 419	8.9	1.15	0.15	0.52	0.02	89.64	91.10
96/0603	8.20	2.10	0.45	0.39	0.02	88.841	91.80
98/2226	9.55	0.35	0.45	0.60	0.02	89.03	90.45
92/0058	8.51	2.80	0.30	0.31	0.05	87.95	91.49
97/0211	8.56	2.80	0.151	0.23	0.10	88.16	91.44
95/0289	9.15	1.40	0.50	0.43	0.03	88.49	90.85
92/0326	9.75	1.75	1.12	0.42	0.08	87.61	90.19
92/1452	9.66	2.80	0.35	0.59	0.02	86.6	90.32
98/0002	8.55	0.36	0.2	0.35	0.12	90.43	91.45
97/4779	8.90	1.05	1.05	0.33	0.02	89.25	91.10
96/1632	9.93	0.35	0.05	0.28	0.08	89.28	90.07
96/0523	8.88	2.80	0.40	0.24	0.03	87.76	91.12
M98/0068	8.28	1.40	0.45	0.42	0.11	89.60	91.72
M98/0028	8.70	0.70	0.45	0.37	0.01	88.94	91.30
96/1089	9.19	0.35	0.15	0.18	0.04	90.09	90.81
95/0166	8.61	2.10	0.05	0.12	0.06	88.8	91.39
92/0057	8.40	1.40	0.15	0.39	0.13	87.09	91.60
96/1314	9.38	1.75	0.30	0.40	0.01	88.14	90.62
97/3200	8.41	1.77	0.60	0.26	0.05	88.92	91.59
98/0040	9.68	1.40	0.10	0.42	0.01	88.02	90.25
TMS30572	7.31	2.45	0.35	0.28	0.14	89.47	92.69
99/2123	8.41	1.05	0.30	0.46	0.15	89.50	91.59
92/0067	10.06	0.70	0.50	0.15	0.20	88.57	89.94
97/0039	8.56	2.80	0.10	0.18	0.23	8.60	91.44
95/0379	9.00	1.40	0.25	0.12	0.11	89.13	91.00
92/0061	8.87	1.40	0.40	0.31	0.03	88.64	91.13
98/0581	8.46	0.35	0.40	0.17	0.17	90.43	91.54
96/1569	8.64	1.85	0.15	0.22	0.01	88.98	91.36

*, **Proximate composition of fufu flours made from CMD-resistant varieties ranging from 1-22 and 23-43 of the 43 varieties under study. Mc: Moisture, CHO: Carbohydrate and DM: Dry Matter content of the flours

RESULTS AND DISCUSSION

Table 1 shows, the secondary data from which cluster observations and similarity levels were examined in order to determine the number of clusters to be used in the subsequent analysis. Using Similarity Level (SL) of the amalgamation steps, Table 2 was obtained. This gave a simple distance measure (Euclidean distance), which can be used to reflect dissimilarity between two patterns. Michalski *et al.* (1983) showed that other similarity measures can be used to characterize the conceptual similarity between patterns. The distance measures with low coefficient are grouped together. A

Table 2: Agglomeration schedule using square euclidean distance measure*, **

Stage	Cluster combined: Cluster 1	Cluster combined: Cluster 2	Coefficient	Stage cluster: Cluster 1	Stage cluster: Cluster 2	Next stage
1	13	43	0.000	0	0	8
2	6	26	0.000	0	0	16
3	14	22	0.017	0	0	28
4	19	20	0.080	0	0	13
5	24	42	0.091	0	0	25
6	3	30	0.107	0	0	25
7	8	37	0.112	0	0	12
8	13	31	0.119	1	0	15
9	21	41	0.214	0	0	19
10	17	34	0.244	0	0	15
11	12	29	0.271	0	0	24
12	8	28	0.317	7	0	23
13	19	39	0.366	4	0	22
14	2	40	0.393	0	0	19
15	13	17	0.486	8	10	30
16	5	06	0.490	0	2	21
17	7	38	0.497	0	0	27
18	1	35	0.508	0	0	28
19	2	21	0.577	14	9	24
20	4	16	0.543	0	0	23
21	5	18	0.587	16	0	27
22	19	27	0.600	13	0	31
23	4	8	0.607	20	12	29
24	2	12	0.747	9	11	26
25	3	24	0.849	6	5	32
26	2	25	0.970	24	0	29
27	5	7	1.069	21	17	28
28	1	14	1.272	18	3	34
29	2	4	1.310	26	23	30
30	2	13	1.352	29	15	36
31	19	33	1.932	22	0	35
32	3	15	2.326	25	0	37
33	10	36	2.741	0	0	39
34	1	23	3.090	28	0	38
35	19	32	3.095	31	0	36
36	2	19	3.573	30	35	39
37	3	9	3.594	32	0	40
38	1	5	4.796	34	27	41
39	2	10	4.831	36	33	40
40	2	3	5.996	39	37	41
41	1	2	6.920	38	40	42
42	1	11	89.120	11	0	0

*, **Agglomeration schedule for fufu flours made from CMD-resistant varieties ranging from 1-22 and 23-42of the 43 varieties under study. The rows are stages of clustering numbered 1 to 43-1

variety of distance measures are in use in the various communities (Anderberg, 1973; Jain and Dubes, 1988; Diday and Simon, 1976).

The agglomeration schedule showed that the varieties could be placed in six groups. The agglomeration schedule shows the amount of error created at each clustering stage when two different objects-cases in the first instance and then clusters of cases are brought together to create a new cluster. A large jump in the value of the error term indicates that two different things have been brought together and that there is a significant typology at that level (David and Roberto, 1998). From the analysis, the coefficient column of the agglomerative schedule is used to establish the clusters. The distance

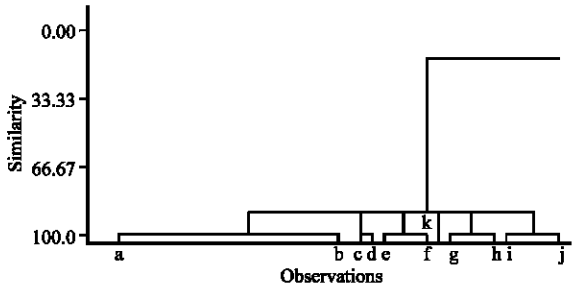

Fig. 1: Dendrogram of fufu flours from the 43 cassava varieties showing 6 cluster groups based on the similarity of their proximate composition. Where cassava varieties that fall into the same cluster groups (observations) are a-b = 13, 43, 31, 17, 34, 8, 37, 28, 4, 16, 12, 29, 21, 41, 2, 40, 25, 19, 20, 39, 27, 33 and 32; c-d = 10 and 36; e-f = 24, 42, 3, 30, 15 and 9; g-h = 7, 38, 6, 26, 5 and 18; i-j = 14, 22, 1, 35 and 23 and k = 11 according to the serial numbers in Table 1

measures with low coefficient, means cases alike, which cluster together. The distance coefficient generated between the 43 cassava varieties ranged from 0.000-89.120 (Table 2).

Cluster observations were executed in this research, subsequently the similarity levels were examined in order to determine the number of clusters to be used in the subsequent analysis. Cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity (Jain *et al*., 1999). In Table 2, the rows are stages of clustering numbered 1 to 43-1. The 42nd stage includes all the cases in a cluster. There are two cluster members for combination at each stage. Stage 1 combines the two cases, which have lowest distance score. The cluster number goes by the lower of the cases or clusters combined, where cases are initially numbered 1-42. For instance at stage 1, cases 13 and 43 are combined resulting in a cluster labeled 13 as in Fig. 1.

A dendrogram of the data, which is also called hierarchical tree diagram or plot, which shows the relative size of the proximity coefficients at which cases were combined. Cases with low distance are close together with a line linking them a short distance from the left of the dendrogram indicating that they are agglomerated into a cluster at a low distance coefficient indicating similarity. In general, the dendrogram shows the pattern of clustering among the varieties with connecting lines further to the right indicating more distance between varieties and clusters.

Figure 1 shows, the dendrogram of the pattern of clustering among the varieties with connecting lines further to the right indicating more distance between varieties and clusters. The dendrogram is a hierarchical tree diagram or plot, which shows the relative size of the proximity coefficient at which cases were combined. Cases with low distances are close together with a line linking them. A short distance from the left of the dendrogram indicates that they are agglomerated into a cluster at a distance coefficient indicating similarity.

The dendrogram can be broken at different levels to yield different clustering of the data. If we draw a horizontal line through the diagram (Fig. 1) at any level on the y-axis (the distance measure, the vertical cluster lines), it intersects indicating clusters, whose members are at least that close to each other. If we draw a horizontal line at some distances, we see that there are 6 clusters. We can see that a case can belong to multiple clusters, depending on where, we draw the line (i.e., how close we require the cluster members to be to each other). Hence, the term hierarchical.

The clustering obtained demonstrated that the proximate composition of each CMD-resistant variety fall into several distinguishable clusters merged as a-b; c-d; e-f; g-h; i-j and k (Fig. 1).

Clusters a-b contained 13, 43, 31, 17, 34, 8, 37, 28, 4, 16, 12, 29, 21, 41, 2, 40, 25, 19, 20, 39, 27, 33 and 32; c-d contained 10 and 36; e-f contained 24, 42, 3, 30, 15 and 9; g-h 7, 38, 6, 26, 5 and 18 while, i-j contained 14, 22, 1, 35 and 23 and k contained only 11 serial numbers of the CMD-resistant varieties (Table 1). The centroid of each of these clusters was determined by computing the mean of the moment vectors of the proximate composition falling into the cluster.

In the emerged clusters, cluster 1 has a total of 23 CMD-resistant varieties based on the similarities of their proximate composition (a-b observations, Fig. 1) and include 98/1565, 96/1569, 95/0166, 96/0603, 97/3200, 98/0505, 99/2123, M98/0068, 97/0162, TME 419, 97/2205, M98/0028, 95/0289, 92/0061, 99/6012, 95/0379, 97/4779, 92/0058, 97/0211, 97/0039, 96/0523, 96/1314 and 92/0057. Cluster 2 has two CMD-resistant varieties (98/2101 and TMS30572) or c-d observations (Fig. 1). Cluster 3 has six CMD-resistant varieties (98/0002, 98/0581, 94/0561, 96/1089, 92/00168 and 99/3037 or e-f observations). Cluster 4 has six CMD varieties in the group including 98/0510, 92/0067, 96/1642, 96/1632, 94/0026 and 98/2226 (g-h observations) while clusters 5 and 6 have five and one CMD varieties stated as 92/0325, 92/0326, 97/4769, 98/0040 and 92/1452 (I-j observation Fig. 1) and 97/4763 (k observation), respectively. Thus, the study has reduced the 43 CMD resistant varieties to 6 cluster groups based on the similarities of their fufu flour proximate composition. This grouping will help the farmers grow only 6 out of the 43 CMD-resistant varieties, one from each group and have almost all the benefits of growing all the forty three varieties at a time.

CONCLUSION

The agglomeration schedule showed that the varieties could be placed in 6 groups. The distance coefficients generated between the 43 cassava varieties ranged from 0.000-89.120. The hierarchical tree diagram or dendrogram showed the relative size of the proximity coefficients at which cases were combined. Cases with low distance are close together with a line linking them, which is a short distance from the left of the dendrogram indicating that they are agglomerated into a cluster at a low distance coefficient indicating similarity.

This study has succeeded in placing the 43 CMD-resistant varieties into cluster groups. The distances measured with low coefficient are grouped together. The dendrogram showed that the forty three varieties could be categorized into 6 cluster groups based on the similarity of their proximate composition. One cluster group has a total of 23 varieties, the 2nd, 2 varieties, while 6 and 5 varieties belonged to 3rd and 4th cluster groups, respectively. Another different 6 varieties belonged to a 5th cluster group, while the 6th cluster group had just one variety.

IMPLICATIONS

The implication of this to the farmer and indeed to the nutritionist is that a variety can be selected from each of the 6 cluster groups for cultivation with the objective of achieving the same nutritional differences in terms of proximate composition without having to examine all the 43 varieties. This is cost effective and time saving, which this experiment was designed to achieve.

REFERENCES

Anderberg, M.R., 1973. Cluster Analysis for applications. Academic Press, Inc., New York.

Crawford, I.M. and R.A. Lomas, 1980. Factory analysis: A Tool for data reduction. Eur. J. Market, 14 (7): 414-421. DOI: 10.1108/EUM0000000004917.

Cunningham, K.M. and J.C. Ogilvie, 1972. Evaluation of hierarchical grouping teachniques: A preliminary study. The Computer J., 15 (3): 209-213. DOI: 10.1093/comjnl/15.3.209.

David, J.F. and M.C. Roberto, 1998. Studies in Multivariate Stratification: Similarity Analysis vs Friedman-Rubin. Joint Statistical Meetings-Section on Survey Research Methods, pp: 996-999.

Diday, E. and J.C. Simon, 1976. Clustering Analysis. Digital Pattern Recognition. In: Fu, K.S. (Ed.). Springer-Verlag, Secaucus, NJ, pp: 47-94.

Etudaiye, H.A., T.U. Nwabueze and L.O. Sanni, 2009. Quality of fufu processed from Cassava Mosaic Disease (CMD) resistant varieties. Afr. J. Food Sci., 3 (3): 61-67. http://www.academicjournals.org/ajfs.

Everitt, B.S., 1993. Cluster Analysis. Edward Arnold, Ltd., London, UK.

FAO, 2002. Food and Agriculture Organization of the United Nations, Agricultural towards 2015/30. Technical Interim Report, April, 2000, Rome. www.fao.org.

Friedman, H.P. and J. Rubin, 1967. On some invariant criteria for grouping data. J. Am. Stat. Assoc., 62: 1159-1178.

IITA, 2005. Growing cassava commercially in Nigeria. Cassava illustration guide book. International Institute of Tropical Agriculture, Ibadan, Nigeria, pp: 21-22.

Jain, A.K. and R.C. Dubes, 1988. Algorithms for Clustering Data. Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ.

Jain, A.K., M.N. Murty and P.J. Flynn, 1999. Data clustering: A review. ACM Computing Surveys, 31 (3): 264-323. http://eprints.iisc.ernet.in/archive/00000273.

Michalski, R., R.E. Stepp and E. Diday, 1983. Automated construction of classifications: Conceptual clustering versus numerical taxonomy. IEEE. Trans. Pattern Anal. Mach. Intell. PAMI-5, 5: 396-409.

Nweke, H., Dustans and L. John, 2002. The Cassava Transformation; African Best Kept Secret. Michigan State University Press, East Lansing, pp: 10.