International Journal of Mathematics and Statistics Invention (IJMSI)
E-ISSN: 2321 – 4767 P-ISSN: 2321 - 4759
www.ijmsi.org Volume 2 Issue 6 || June. 2014 || PP-44-51

Cluster Analysis of Some Selected Eye Diseases: A case study of Gombe State, Nigeria

[1,]Yakubu Musa [2,]A. B. Zoramawa
[1,]*Department of Mathematics, Statistics Unit, Usmanu Danfodiyo University, Sokoto*
[2,]*Department of Mathematics, Statistics Unit, Usmanu Danfodiyo University, Sokoto, Nigeria.*

ABSTRACT : *In this paper, data of some selected eye diseases were collected from federal medical centre in Gombe State, Nigeria for the period of 1997 to 2009. The Analysis were carryout using multivariate technique with particular reference to hierarchical clustering techniques, such as Single Linkage, Complete Linkage, simple average, and Centroid Methods to partition the set of diseases into groups such that the diseases with similar degree of prevalence were identified. The result of the cluster formation shows that disorder of conjunctiva is more prevalent from almost all of the five method employed, followed by cataract, then Glaucoma. Using Goodness of fit, the study concludes that the use of simple linkage (weighted pair-group method) of clustering methods is the best that model the mechanism, which provides a suitable tool for assessing the level of infections of the disease, this procedure become more useful in medical research to classify diseases.*

KEYWORDS: *Eye diseases, hierarchical clustering, Gombe*

I. INTRODUCTION

Many developed countries used statistical data to improve in their economics, industries, education and medical sector. It is in realization of this that government all over the world are striving to cater for the well being of their population. Health is very important area of investment in human endeavour. A healthy population bring about a healthy economic, however critical analysis of health becomes difficult for lack of sufficient data, it is difficult to sufficiently facilitate the execution of health plan without statistical data needed for such plan.

Three major problems are contributing to a growing "health crisis" which has already increased the burden of the health to an in supportable level, these problems are:
[1] the magnitude and diversity of health hazard associated with development
[2] the cost of treating these diseases cause by industrialization and urbanization,
[3] and the need for macro economic adjustment which has resulted in major cuts in the
[4] health budged in many developing countries.

Despite all odds associate with health data, this research work will focus attention on the analysis of some reported cases of eye diseases in Nigeria taking federal medical centre Gombe as a case study.Due to the uncountable number of eye diseases case in the federal medical centre Gombe, this research will attach more emphasis on the 5 most prevailing types of eye diseases in Gombe and its environs that can be treated at federal medical centre (F.M.C) Gombe, these are: cataract, glaucoma, retina detachment, disorder conjunctiva, and blindness. Infection differs from other diseases in a number of aspects. The most important is that it is caused by living microorganisms which can usually be identified, thus establishing the aetiology early in the illness. Many of these organisms, including all bacteria, are sensitive to antibiotics and most infections are potentially curable, unlike many non-infectious diseases which are degenerative and frequently become chronic. Communicability is another factor which differentiates infectious from non-infectious diseases. Transmission of pathogenic organisms to other people, directly or indirectly, may lead to an epidemic. Finally many infections are preventable by hygienic measures, by vaccines or by the judicious use of drugs (chemoprophylaxis) *(Davidson, 2006).*For these reasons, therefore, statisticians and social scientists used different scientific methods to analyze the cultural and behavioural aspects of the infectious diseases as well as their impact on families, communities and nations in general.One of the most commonly used scientific methods is multivariate analysis. Multivariate analysis consists of a collection of methods that can be used when several measurements are made on each individual or object in one or more samples. We will refer to the measurements as variables and to the individuals or objects as units (research units, sampling units, or experimental units) or observations. In practice, multivariate data sets are common, although they are not always analyzed as such.

But the exclusive use of univariate procedures with such data is no longer excusable, given the availability of multivariate techniques and inexpensive computing power to carry them out.Morrison (1990) opined that multivariate statistical analysis is concerned with data collected on several dimensions of the same individual. Such observations are common in the social, behavioural life and medical science. It therefore, helps the researcher to summarize the data and reduce the number of variable necessary to describe it.

Clustering technique (one of the multivariate approach), have been applied to a wide variety of research problems. Hartigan (1972) provides an excellent summary of the many published studies reporting the results of cluster analysis. For example, in the field of medicine, clustering of disease cures for diseases, or symptoms of diseases have lead to very useful taxonomies. In the field of psychiatry, the correct diagnosis of cluster of symptoms such as paranoia, schizophrenia is essential for successful therapy. In archaeology, researchers have attempted to establish taxonomies of stone tools, funeral objects, e.t.c. by applying cluster analysis techniques.Cluster analysis seeks to partition a set of individuals into some form of natural groupings, if any. It is one tool of exploratory data analysis that attempt to assess the interaction among patterns by organizing the patterns into groups or cluster, such that patterns within cluster are more similar to each other than are pattern belonging to different clusters *(Hartigan, 1972).*One of the most challenging tasks to public health in Nigeria and Africa in general, is the control of common infectious diseases. Most of these diseases have already been eliminated in Europe and the Americas.

The problem in Nigeria especially, lies mainly in the behaviour or lack-luster attitude of the people towards public health. The environment is littered with excretes (a medium for cholera), carcasses (medium for viral/bacterial infections), contaminated ponds, stagnant water and blocked drainages (breeding medium for mosquitoes) and polythene bags (item blocking soil pores /water passages). In addition the country is yet to have a full working system of hygienic drinking water etc.Therefore, to achieve full and effective public health status, there is the need to study the prevalence and intensity of the infectious diseases with a view to helping the authorities concerned put in place sound policies and programmes towards achieving healthy population.
Medical, biological, and industrial experiments all benefit from statistical methods and theories. The results of all of them serve as predictors of future performance, though reliability varies.The aim of this paper is to apply some of the multivariate techniques with particular reference to hierarchical clustering techniques, such as Single Linkage, Complete Linkage, simple average, Centroid and Ward's Methods to partition some set of diseases into groups, with the objectives of determining which of these eye diseases is more prevalent in the study area and to make comparisons among the cluster analysis techniques applied to understand the best method among the different methods employed.

Study Area Location : Gombe State is located between latitude 9'30' and 12'30'N and longitudes 8'45' and 11 '45'E of the Greenwich Meridian. It lies within the Northeast region of Nigeria and occupies a total land area of about 20,265sq.Km. The state had, by 2006, an estimated population of 1,820,415 inhabitants.Gombe State, "Jewel in the Savannah," was created on 1st October, 1996 by the General Sani Abacha administration. Its creation was a fulfilment of the aspirations of the people who, for long, had been yearning for a state of their own out of the then Bauchi State. The genesis of the struggle dates back to 1979. Since its creation, the state has been growing fast, blessed with abundant physical, human and economic resources. One of the criteria for an area to qualify for State creation is economic viability. There is no doubt that Gombe fulfilled that criterion. Gombe State comprised eleven Local Government Areas (LGAs) However, some of the LGAs, notably Akko, Dukku and Gombe have been in existence since 1976, as part of the then Bauchi State.Gombe town is the administrative capital of the state, as it has been for the Emirate, the seat of the Emir. Its population of about 200,000 makes it the largest centre in the state. It is also the commercial and zonal services headquarters in Nigeria for many Federal Government establishments, including NNPC, NEPA, NITEL Territorial Office, NIPOST Territorial Office and Nigerian Railway Corporation

Definition of eye diseases
Cataract : Cataract is clouding of the lens of the eye which impedes the passage of light.

Glaucoma: Glaucoma can be regarded as a group of diseases that have as a common end point a characteristic optic neuropathy which is determined by both structural change and functional deficit.

Retina Detachment : Retinal detachment is a disorder of the eye in which the retina peels away from its underlying layer of support tissue.

Blindness and Low Vision: Blindness is the condition of lacking visual perception due to physiological or neurological factors.

Disorder of Conjunctiva: Disorders of the conjunctiva and cornea are a common source of eye complaints. The surface of the eye is exposed to various external influences and is especially susceptible to trauma, infections, chemical irritation, allergic reactions and dryness.

II. MATERIAL AND METHODS

Data

The data collected and used in this paper were the number of patient treated with five different eye diseases. The data were collected from Federal Medical centre (FMC) Gombe, from 1997 to 2009 annually.

Cluster Algorithm

There are two types of clustering algorithm: hierarchical and non-hierarchical (partitioning) algorithm.

Hierarchical Algorithms (Agglomerative Techniques)

Hierarchical methods and other clustering algorithm represent an attempt to find "good" cluster in the data using a c computationally efficient technique (Rencher, 2002). Obviously time constraint make it impossible to determine the best grouping of similar objects from the list all possible structure, especially a larger one (Richard and Dean, 2001). According to Duran and Odell (1974), the number of ways of partitioning a set of n items into g cluster is given by

$$N(n,g)=\frac{1}{g!}\sum_{k=1}^{g}\binom{g}{k}(-1)^{g-k}k^{n}$$

This can be approximated by $\frac{g^n}{g!}$, which is large even for moderate value of n and g.

Procedure for calculating hierarchical algorithms;The algorithm used by all the following five clustering methods is out line as follows. Let the distance between clusters i and j be represented as d_{ij} and let cluster i contain n_1 object. Let D represent the set of all remaining d_{ij}. Suppose there are N objects to cluster.

[1] Find the smallest element d_{ij} remaining in D.

[2] Merge clusters i and j into a single new cluster, k.

[3] Calculate a new set of distance d_{km} using the following distance formula; $d_{km}=\alpha_i d_{im}+\alpha_j d_{jm}+\beta d_{ij}+\gamma\left|d_{im}-d_{jm}\right|$. Here m represents any cluster other than k. These new distance replace d_{im} and d_{jm} in D. Also let $n_k=n_i+n_j$. Note that the five algorithm available represent five choices for α_i,α_j,β and γ.

Repeat steps1-3 until D contains a single group made up off all objects; this will require N-1 iteration.
Five different methods for creating clusters using hierarchical cluster analysis are available, these include: - single linkage, complete linkage, average linkage, centroid method and the ward's method.

Hierarchical Methods for Creating Clusters

Single Linkage: this is also known as nearest neighbour clustering, this is one of the oldest and most famous of the hierarchical techniques. The distance between two groups is defined as the distance between their two closer members.

At each step in the single linkage method, the distance is found for every pair of clusters, and the two clusters with smallest distance are merged. The number of clusters is therefore reduced by 1. After two clusters are merged, the procedure is repeated for the next step: the distances between all pairs of clusters are calculated again, and the pair with minimum distance is merged into a single cluster. *(Rencher, 2002).* The results are displayed graphically using a tree diagram, also known as a dendrogram.

The coefficients of the distance equation are; $\alpha_i=\alpha_j=0.5,\beta=0,\gamma=-0.5.$

Complete Linkage: also known as furthest neighbour or maximum method, this method defined the distance between two groups as the distance between their two farthest-apart members. This method usually yields clusters that are well separated and compact.

The coefficients of the distance equation are: $\alpha_i = \alpha_j = 0.5, \beta = 0, \gamma = 0.5$

Simple Average: also called the weighted pair-group method, this algorithm defined the distance between groups as the average distance between each of the members, weighted so that the two groups have an equal influence on the final result.

The coefficients of the distance equation are: $\alpha_i = \alpha_j = 0.5, \beta = 0, \gamma = 0$

Centroid Method: also known as un-weighted pair-group centroid method, this method defined the distance between two groups as the distance between their centroids (centre of gravity or vector average). The method should only be used with Euclidean distance.

The coefficients of the distance equation are: $\alpha_i = \frac{n_i}{n_k}, \alpha_j = \frac{n_j}{n_k}, \beta = -\alpha_i \alpha_j, \gamma = 0$.

A restriction of employing this method is that it requires metric data.

Other Additional Techniques.

Dendrogram: this is a pictorial representation of the clustering process that identifies how observations are combined with one another to form the extracted clusters.

A dendrogram that clearly differentiates groups of objects will have small distance in the far branches of the tree and large differences in near branches.

Goodness-Of-Fit: Given the large number of techniques, it is often difficult to decide which is best. One criterion that has become popular is to use the result that has largest *cophenetic correlation coefficient*. This is the correlation between the original distances and those that result from the cluster configuration. Values above 0.75 are felt to be good. The Group Average method appears to produce high values of this statistic. This may be one reason that it is so popular.

A second measure of goodness of fit called *delta* is described in Mather (1976). These statistics measure degree of distortion rather than degree of resemblance (as with the cophenetic correlation).

Data Analysis :

Following are the data analysis of some eye diseases in Gombe from 1997 to 2009 a case study of federal medical center Gombe.

Single Linkage (Nearest Neighbor).

Hierarchical Clustering Report for Single Linkage (Nearest Neighbor)

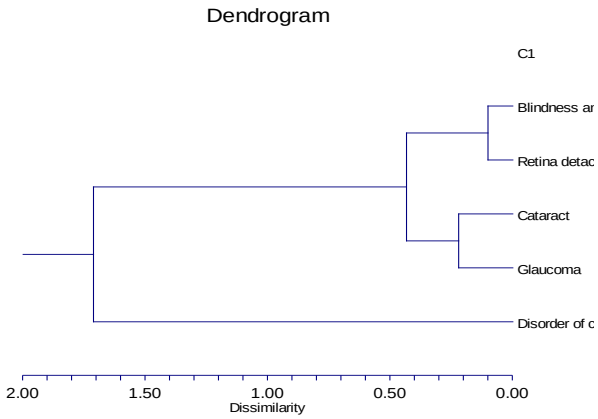
Cluster Detail Section

Row	Cluster	C1
2	1	Glaucoma
3	1	Retina detachme
4	1	Cataract
5	1	Blindness and l
1		Disorder of con

Linkage Section

Link	Number Clusters	Distance Value	Distance Bar	Rows Linked
4	1	1.712639	|IIIIIIIIIIIIIIIIIIIIIIIIIIIII	1,2,4,3,5
3	2	0.434305	|IIIIIIII	2,4,3,5
2	3	0.221305	|IIII	2,4
1	4	0.101528	|II	3,5

Cophenetic Correlation	0.968576
Delta(0.5)	0.255843
Delta(1.0)	0.311265



The dendrogram visually displays a particular cluster configuration. Rows that are close together (have small dissimilarity) has been linked near the right side of the plot. For example, we notice that Glaucoma and cataract are very similar. Rows that link up near the left side are very different. For example, disorder of conjunctiva appears to be quite different from any of the other eye diseases.

Complete Linkage (Furthest Neighbor) method

Hierarchical Clustering Report for Complete Linkage (Furthest Neighbor) method.

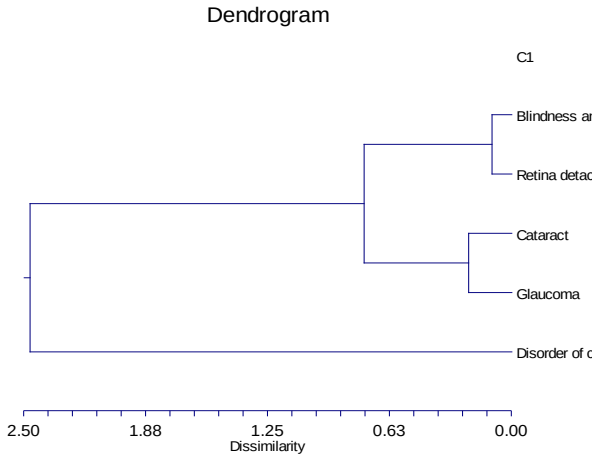
Cluster Detail Section

Row	Cluster	C1
2	1	Glaucoma
3	1	Retina detachme
4	1	Cataract
5	1	Blindness and l
1		Disorder of con

Linkage Section

Link	Number Clusters	Distance Value	Distance Bar	Rows Linked
4	1	2.469776	|IIIIIIIIIIIIIIIIIIIIIIIIIIIII	1,2,4,3,5
3	2	0.757138	|IIIIIIIII	2,4,3,5
2	3	0.221305	|III	2,4
1	4	0.101528	|I	3,5

Cophenetic Correlation	0.968873
Delta(0.5)	0.154202
Delta(1.0)	0.196619



Simple Average (Weighted Pair-Group) method.

Hierarchical Clustering Report for Simple Average (Weighted Pair-Group) method.

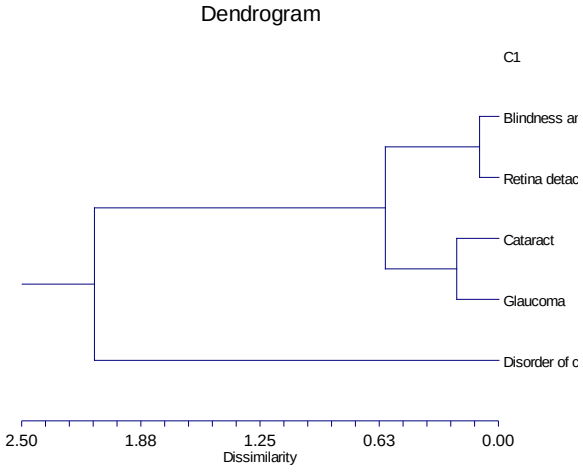
Cluster Detail Section

Row	Cluster	C1
2	1	Glaucoma
3	1	Retina detachme
4	1	Cataract
5	1	Blindness and l
1		Disorder of con

Linkage Section

Link	Number Clusters	Distance Value	Distance Bar	Rows Linked
4	1	2.121152	|IIIIIIIIIIIIIIIIIIIIIIIIIIIII	1,2,4,3,5
3	2	0.595721	|IIIIIIII	2,4,3,5
2	3	0.221305	|III	2,4
1	4	0.101528	|I	3,5

Cophenetic Correlation	0.969538
Delta(0.5)	0.146024
Delta(1.0)	0.150955



Centroid (Un weighted Pair-Group Centroid) method.

Hierarchical Clustering Report for Centroid (Un weighted Pair-Group Centroid) method.

Cluster Detail Section

Row	Cluster	C1
2	1	Glaucoma
3	1	Retina detachme
4	1	Cataract
5	1	Blindness and l
1		Disorder of con

Linkage Section

Link	Number Clusters	Distance Value	Distance Bar	Rows Linked
4	1	1.952044	|IIIIIIIIIIIIIIIIIIIIIIIIIIIII	1,2,4,3,5

3	2	0.515013	|IIIIIIII	2,4,3,5
2	3	0.221305	|III	2,4
1	4	0.101528	|II	3,5

Cophenetic Correlation	0.969267
Delta(0.5)	0.164428
Delta(1.0)	0.188965

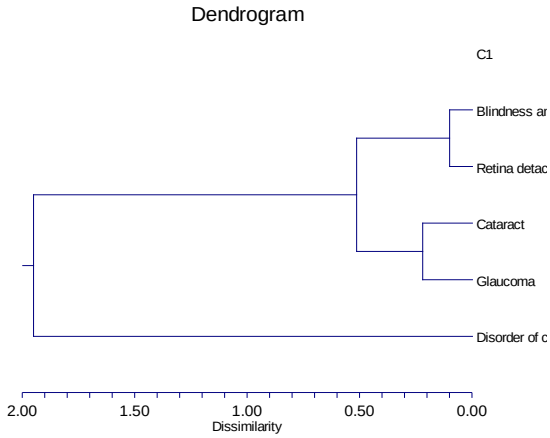


Table 3.1: Summary of the Cluster Formation Analysis

Clusters	**Methods for cluster formation**			
	Single Linkage	Complete Linkage	Simple Linkage	Centroid Method
4	1,2,4,3,5	1,2,4,3,5	1,2,4,3,5	1,2,4,3,5
3	2,4,3,5	2,4,3,5	2,4,3,5	2,4,3,5
2	2,4	2,4	2,4	2,4
1	3,5	3,5	3,5	3,5

1Disorder of conjunctiva, 2. Glaucoma, 3. Retinal detachments, 4. Cataract,
5. Blindness,

The Table 3.1 presents the summary of cluster formation analysis for the data, and it indicates that Disorder of conjunctiva is more prevalent disease in the study area, irrespective of the cluster formation method employed. It is closely followed by cataract, irrespective of the method employed.

Disorder of conjunctiva and cataract were followed by Glaucoma, then Blindness and low vision and finally Retina detachment. The five diseases from a single cluster of closely related diseases are statistically significant. The cluster detail section report displays the cluster number associated with each row. The report is sorted by row number within cluster number. The cluster numbers of rows that cannot be classified are left blank. The cluster configuration depends on the Cluster Cut-off value that was used.The Linkage section report displays the subgroup that is formed at each fusion that took place during the cluster analysis. The links are displayed in reverse order so that we can quickly determine an appropriate number of clusters to use. It displays the distance level at which the fusion took place. It will let us precisely determine the best value of the Cluster Cutoff value. For example, looking down the Distance Value column of the report, one can see that the cutoff value that had been used in the case of single linkage (the default value is 1.0) occurs between Links 3 and 4. Hence, the cutoff value of 1.0 results in three clusters.The cophenetic correlation section and the two delta goodness of fit statistics that are reported at the bottom of the report are the values that will let us compare the fit of various cluster configurations

Table 3.2: summary of the cophenetic correlation and delta values.

Linkages.	Cophenetic values.	Delta 0.5	Delta 1.
Single linkage	0.96	0.25	0.31
Complete linkage	0.96	0.15	0.19
Simple average	0.96	0.14	0.15
Centroid	0.96	0.16	0.18
Wards method.	0.96	0.33	0.34

From our previous discussion, one can see that all the five different clustering method meet the requirement for the clustering to be considered, since the cophenetic correlation values for all the methods is greater than 0.75. But for the goodness of fit deltas, only simple linkage method meet the requirement to fit the data better since it is the method that has the smallest delta values.The dendrogram reports above, visually displays a particular cluster configuration. Rows that are close together (have small dissimilarity) will be linked near the right side of the plot. For example, we notice the Cataract and Glaucoma are very similar likewise blindness and retina detachment are also similar.

III. CONCLUSION

In this paper, analysis was carried out using a multivariate approach of five different cluster techniques, within the five techniques used; one can see that all the five different clustering method meet the requirement for the clustering to be considered, since the cophenetic correlation values for all the methods is greater than 0.75. But for the goodness of fit deltas, only simple linkage method meet the requirement to fit the data better since it is the method that has the smallest delta values. This means that simple linkage method is the best for analyzing the eye diseases data, for instance using the simple linkage techniques, and we conclude that Glaucoma and Cataract share the same patterns or features in effecting the population in the area under study. Disorder of conjunctiva being the most prevalent eye disease followed by cataract and Glaucoma then blindness and low vision lastly retina detachment in the study area location.Therefore there is the need for government or authority concerned to put in place sound programmes for the eradication of such diseases and provision of more medical services and facilities.

REFERENCES

[1] Aldernderfer, M. S. & Blashfield R. K. (1984). Cluster Analysis. Sage Publication, Newbury park.
[2] Everitt, B.S. (1974), Cluster Analysis. Heinemann Educational Books Limited U.K.
[3] Gulumbe, S. U., Bakar, A.B. and Dikko, H.G. (2008), Classification of some HIV/AIDS Variables, a multivariate approach. Research Journal of Science 15, 24 – 30
[4] Hardle, W. and Simar, L. (2007), Applied Multivariate Statistical Analysis. Spring Berlin Heidelberg, New York.
[5] Hartigan, J.A. (1972), Direct clustering of a data matrix. Journal of the American Statistical Association. 67, 123 – 129.
[6] Joshua, B. P., Jerome, C. and Peter, S. A. (2002), Cluster analysis of spatial patterns in Malaysian tree species. American Naturalist Journal 160, 629 – 644.
[7] Legal Notice on publication (2006), Census final results. National Population Commission.
[8] Morrisson, D.F. (1990), Multivariate Statistical Methods. McGraw-Hill Book Company, New York.
"Revision of the International Classification of Diseases (ICD)". World Health Organization. Retrieved 29 October 2010
[9] Rencher, A. C. (2002), Method of Multivariate Analysis 2nd edition. John Wiley & Sons Inc, Third Avenue, New York.
[10] Roberta, B. N. & el tal (2005), Cluster analysis of bacterial Vaginosis associated microflora and pelvic inflammatory disease. American Journal of Epidemiology 162, 1 – 6 .