NJBAS

Available online at http://www.ajol.info/index.php/njbas/index
Nigerian Journal of Basic and Applied Science (2011), 19 (1): 31-42

ISSN 0794-5698

Monetering of Infectious Diseases in Katsina and Daura Zones of Katsina State: A Clustering Analysis

[1]U. Dauda, [2]S.U. Gulumbe, *[2]M. Yakubu and [1]L.K. Ibrahim
[1]Department of Mathematics and Computer Science, Umaru Musa Yar'aduwa University, Katsina.
[2]Department of mathematics Usmanu Danfodiyo University, Sokoto Nigeria
[*Corresponding Author: ykmtm2000@yahoo.com]

ABSTRACT: In this paper, data of infectious diseases were collected from the two senatorial zones of Katsina state, and analyzed using cluster analysis, a multivariate technique. This necessitated a partition of 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 Malaria is more prevalent in all of the two zones, followed by Cholera and Typhoid fever using the Single Linkage and Centroid methods. The Complete Linkage and Ward methods showed that Malaria is the most prevalent followed by Typhoid fever and Cholera in Katsina zone, while in Daura zone Typhoid fever is more prevalent followed by Malaria and Cholera. The number of clusters tends to vary from one zone to another. This is achieved by using Chi-square test for independence. The study concludes that the use of clustering methods provides a suitable tool for assessing the level of infections of the disease.
Keywords: Cluster analysis, Infectious diseases, Malaria, Cholera and Typhoid

INTRODUCTION

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 America. The problem in Nigeria especially, lies mainly in the behavior or lack-luster attitude of the people towards public health. The environment is littered with excrete (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.

Diseases affecting humans are caused by infection. Such as leprosy, chickenpox and typhoid fever (Bloom, 1963). The aetiology of some of these diseases is induced by environmental factors. 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 behavioral 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.

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

attempts 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).

Gulumbe *et al*. (2008) applied hierarchical clustering techniques to partition the set of variables into groups, such that those are similar with respect to HIV/AIDS. Infections were identified and two main clusters were observed. The implication of the cluster formation shows that HIV/AIDS infection is more prevalent among married women as in single and wardôs linkage methods. It also shows that the disease affect mostly the working class aged from 15 to 39 as grouped by complete linkage method. The relationship between the various methods used and clusters formed with respect to the variable grouped were found to be consistent using chi-square test for independence.

Solovyov *et al*. (2009) applied cluster analysis for the origins of the new influenza A (H1N1 virus). They reported that A (H1N1) virus was the reassortment of at least two swine influenza viruses from North America (in light blue) and Eurasia (in dark blue). Blanchette and Marks (2000) applied single linkage method to gene expression analysis of cancer patients. The results gave a comprehensive understanding of the mostly subtle difference in gene expression of different tumor types, which is crucial for elucidating the molecular mechanisms of cancer as well as for the successful treatment of the disease.

Mclaren (1999) applied single and complete linkage methods for the screening of objects for blood-related diseases based on bivariate histogram, measuring red-blood cell volume and haemoglobin content. The research work led to some promising discoveries in haematology, the study of blood and blood related disease.

A general question often faced by researchers in many areas of inquiries, is how to organize observed data into meaningful structures. The ability to achieve this is essential, if one is to make sense out of the tremendous diversity of organisms, diseases e.t.c. One of the most commonly used terms for techniques, which seek to separate data into constituent groups is cluster analysis. It is a collection of statistical methods that can be used to assign cases or individuals to a group so that group members will share certain common properties.

In this paper therefore, the use of cluster analysis (Single, Complete, Centroid and Ward Methods) to monitor the pattern of infectious disease outbreak in Katsina state will be explored. The objectives of this paper are to: (i) classify the prevalence of diseases according to the zones. (ii) Identify the most prevalent disease in each zone. (iii) Compare the analyzed results from the two different zones.

Data: The data for this paper covered the period of 36 months from January, 2006 to December, 2008. The data was obtained from Babbar Ruga General Hospital Katsina and General Hospitals of the two different zones of Katsina State: Daura General Hospital and Katsina General Hospital.

Study Area

Location: Katsina State, located at the extreme northern margin of Nigeria, covers a total area of about 23,938sqkm (3,370sq) with a total population of 5,801,584 people and lies between latitude $11^0$08ôN and $13^0$22ôN ($13^0$00ôN-$13^0$25ôN) and longitude $6^0$52ôE ($7^0$37ôE and $8^0$00ôE), with thirty four (34) local governments. The local governments are divided into three (3) senatorial zones according to their geographical locations. The zones are as follows:

i. **Katsina Zone:** Batagarawa (189,059), Batsari (207,874), Charanchi (136,989), Danmusa (113,190), Jibia (167,435), Kaita (182,405), Katsina (318,132), Kurfi (116,700), Rimi (154,092), Safana (185,207), Dutsinma (169,829), going by 2006 census (FGN, 2007)

ii. **Daura Zone:** Baure (202,941), Bindawa (151,002), Daura (224,884), Dutsi (120,902), Ingawa (169,148), Kankia (151,397), Kusada (98,348), Maiôadua (201,800), Mani (176,301), Mashi (171,070), Sandamu (136,944), Zango (156,052), going by 2006 census (FGN, 2007).

Climate: The climate is hot and dry for most of the year, maximum day temperature of about 38^0C in the month of March, April and May are common and the minimum temperature is about

22^0C in the month of December and January and Rainfall annual average of 780mm.

MATERIALS AND METHODS

The method used for the classification of the diseases according to the two senatorial zones of Katsina State is hierarchical clustering techniques. The emphasis is on Single Linkage Method, Complete Linkage Method, Centroid Method and Wardôs Method. These methods are generally suitable for searching of natural clusters and they perform reasonably well when clusters are clearly separated (Everitt, 1974). The four linkage methods were used, as this will help to prevent misleading results being accepted. However differences in the linkage methods are due to differences in defining distance (similarity) between groups for each of the methods (Everitt, 1974).

Measures of Proximity: Since cluster analysis attempts to identify the observation vectors that are similar and group them into clusters, many techniques use an index of proximity between each pair of observations. A convenient measure of proximity is the distance between two observations. Since a distance increases as two units become further apart, distance is actually a measure of dissimilarity.

A common distance function is the Euclidean distance between two vectors

$X = \left(x_1, x_2, ..., x_p\right)'$ and $Y = \left(y_1, y_2, ..., y_p\right)'$,

defined as

$$d(x,y) = \sqrt{(x-y)'(x-y)} = \sqrt{\sum_{j=1}^{p}\left(x_j - y_j\right)^2}. \quad (1)$$

To adjust for differing variances and covariances among the p variables, we could use the statistical distance

$$d(x,y) = \sqrt{(x-y)' s^{-1}(x-y)} \quad (2)$$

Where S is the sample covariance matrix. After the clusters are formed, S could be computed as the pooled within-cluster covariance matrix.

Hierarchical Algorithms (Agglomerative Techniques): The hierarchical attempt to find ñgoodò clusters in the data using a computationally efficient technique. The method is also used quite frequently in practice; the algorithm consists of the following steps:

(i) Construct the finest partition.
(ii) Compute the distance matrix D.
(iii) Find the two clusters with the closest distance.
(iv) Put those two clusters into one cluster. (v) Compute the distance between the new groups and obtain a reduced distance matrix D.

UNTIL all clusters are agglomerated into X. (Hardle and Simar, 2007).

Single Linkage (Nearest Neighbor): In the single linkage method, the distance between two clusters A and B is defined as the minimum distance between a point in A and a point in B:

$D(A,B) = \min\{ d\left(y_i, y_j\right), \text{for } y_i \text{ in } A \text{ and } y_j \text{ in } B \}, \quad (3)$

where $d\left(y_i, y_j\right)$ is the Euclidean distance in (1) or some other distance between the vectors y_i and y_j. This approach is also called the nearest neighbor method.

At each step in the single linkage method, the distance (3.3) 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).

Complete Linkage (Farthest Neighbor): In the complete linkage approach, also called the farthest neighbor method, the distance between two clusters A and B is defined as the maximum distance between a point in A and a point in B:

$D(A,B) = \text{Max } \{ d\left(y_i, y_j\right), \text{for } y_i \text{ in } A \text{ and } y_j \text{ in } B \}, \quad (4)$

At each step, the distance (3.4) is found for every pair of clusters, and the two clusters with the smallest distance are merged (Rencher, 2002).

Centroid Method: In the centroid method, the distance between two clusters A and B is defined

as the Euclidean distance between the mean vectors (often called centroids) of the two clusters: $D(A,B) = d(\bar{y}_A, \bar{y}_B)$, (5)

where $\overline{y_A}$ and $\bar{y}_B$ are the mean vectors for the observation vectors in A and the observation vectors in B, respectively, and $d(\bar{y}_{A,}\bar{y}_B)$ is defined in (3.1). We define $\bar{y}_A$ and $\bar{y}_B$ in the usual way, that is, $\bar{y}_A = \sum_{i=1}^{nA} \frac{yi}{nA}$. The two clusters with the smallest distance between centroids are merged at each step. After two clusters A and B are joined, the centroid of the new cluster AB is given by the weighted average (Rencher, 2002) as:

$$\bar{y}_{AB} = \frac{n_A \bar{y}_A + n_B \bar{y}_B}{n_A + n_B}. \quad (6).$$

WARD'S Method: Wardôs method, also called the incremental sum of squares method, uses the within cluster (squared) distances and the between-cluster (squared) distances. If AB is the cluster obtained by combining clusters A and B, then the sum of within-cluster distances (of the items from the cluster mean vectors) are:

$$SSE_A = \sum_{i=1}^{n_A} (y_i - \bar{y}_A)' (y_i - \bar{y}_A), \quad (7)$$

$$SSE_B = \sum_{i=1}^{n_B} (y_i - \bar{y}_B)' (y_i - \bar{y}_B), \quad (8)$$

$$SSE_{AB} = \sum_{i=1}^{n_{AB}} (y_i - \bar{y}_{AB})' (y_i - \bar{y}_{AB}), \quad (3.9)$$

Where $\bar{y}_{AB} = \frac{(n_A \bar{y}_A + n_B \bar{y}_B)}{(n_A + n_B)}$, as in (3.6), and n_A, n_B, and $n_{AB} = n_A + n_B$ are the numbers of points in A, B, and AB, respectively. Since these sums of distances are equivalent to within-cluster sums of squares, they are denoted by: SSE_A, SSE_B and SSE_{AB} (Rencher, 2002).

Wardôs method joins the two clusters A and B that minimize the increase in SSE, defined as

$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B). \quad (10)$$

Test of Independence: The hypothesis we wish to test for is whether the number of clusters formed by different methods varies from one zone to the other, using test of independence.

The test statistic under the null hypothesis is $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(0_{ij} - e_{ij})^2}{e_{ij}}$ which is distributed approximately as X^2_{α} (v) variate with (r-1)(c-1) degrees of freedom O_{ij} is the observed frequency while e_{ij} is the expected frequency in the cell. We state the null hypothesis as:

H_O: Number of clusters formed by different methods does not vary from one zone to another

H_1: Number of clusters formed by different methods varies from one zone to another.

Critical region: Accept H_o if the calculated value is greater than the tabulated value at $\propto$ % level of significance and degree of freedom (r-1)(c-1), otherwise reject.

Data Analysis and Interpretation: In this paper, eight notifiable infectious diseases with their occurrences in various zones of Katsina State were studied, they include: Leprosy, Tuberculosis, Chicken pox, Typhoid fever, Malaria, Cholera, Tetanus and Measles.

For easy representation, let Infectious Diseases under Study be represented as: Leprosy (= 1), Tuberculosis (= 2), Chicken pox (= 3), Typhoid fever (= 4), Malaria (=5), Cholera (=6), Tetanus (=7), Measles (=8).

We used SPSS 15.0 as our analysis tool.

Analysis and Interpretation of Data from Katsina Zone

Table 1 gives the summary of valid and missing cases disease. Six diseases were observed and two diseases Leprosy and Tetanus were not considered because of their high Euclidean distance. The results in Table 1 Case Analysis of Katsina Zone data using Single Linkage Method are the same result by using the remaining methods: Complete Linkage Method, Centroid Method and Ward Method. In all the methods six diseases were observed and two diseases Leprosy and Tetanus

were not considered because of their high Euclidean distance.

Table 1: Case Analysis of Katsina Zone data using Single Linkage Method.

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
6	75.0	2	25.0	8	100.0

Proximity Matrix Analysis

The proximity matrix analysis for Katsina zone data using Single Linkage method, Complete Linkage Method, Centroid Method and Ward Method shows the relationship between any two of the diseases. The result in Table 2 is the same results for the remaining three methods. The results for all the methods show the distance between the diseases. Table 3 presents the agglomeration schedule that shows the stages involved in forming the clusters based on the Euclidean distances between the diseases.

Table 4 gives the following clusters result: 1st Cluster: Malaria; 2nd Cluster: Malaria and Cholera; 3rd Cluster: Malaria, Cholera and Typhoid Fever; 4th Cluster: Malaria, Cholera, Typhoid Fever and Chickenpox; 5th Cluster: Malaria, Cholera, Typhoid Fever, Chickenpox and Measles. Table 5 presents the agglomeration schedule that shows the stages involved in forming the clusters based on the Euclidean distances between the diseases.

By using the complete linkage method , the analysis result gives the following clusters : 1st Cluster: Malaria; 2nd Cluster: Malaria and Typhoid Fever; 3rd Cluster: Malaria , Typhoid Fever and Cholera ; 4th Cluster: Malaria, Typhoid Fever , Cholera and Measles; 5th Cluster: Malaria, Typhoid Fever, Cholera, Measles and Chickenpox (Table 6). Table 7 presents the agglomeration schedule that shows the stages involved in forming the clusters based on the Euclidean distances between the diseases.

Table 2: Proximity matrix of Katsina zone data using Single Linkage method

Case	Squared Euclidean Distance					
	2	3	4	5	6	8
2:TUBERCULOSIS	.000	103183.000	598843.000	2690039.000	925488.000	164970.000
3:CHICKENPOX	103183.000	.000	764432.000	2974774.000	1270541.000	37083.000
4:TYPHOIDFEVER	598843.000	764432.000	.000	1325940.000	783045.000	836951.000
5:MALARIA	2690039.000	2974774.000	1325940.000	.000	2004733.000	2923843.000
6:CHOLERA	925488.000	1270541.000	783045.000	2004733.000	.000	1408274.000
8:MEASLES	164970.000	37083.000	836951.000	2923843.000	1408274.000	.000

2 Tuberculosis, 3 Chickenpox, 4 Typhoid fever, 5 Malaria, 6 Cholera, 8 Measles

Table 3: Agglomeration Schedule for Katsina zone data using Single Linkage method.

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
1	3	8	37083.000	0	0	2
2	2	3	103183.000	0	1	3
3	2	4	598843.000	2	0	4
4	2	6	783045.000	3	0	5
5	2	5	1325940.000	4	0	0

Table 4: Cluster formation by cases for Katsina zone data using Single Linkage method.

Number of clusters	Case										
	5		6		4		8		3		2
1	X	X	X	X	X	X	X	X	X	X	X
2	X		X	X	X	X	X	X	X	X	X
3	X		X		X	X	X	X	X	X	X
4	X		X		X		X	X	X	X	X
5	X		X		X		X	X	X		X

2 Tuberculosis, 3 Chickenpox, 4 Typhoid fever, 5 Malaria, 6 Cholera, 8 Measles

Table 5: Agglomeration Schedule for Katsina zone data using Complete Linkage method.

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
1	3	8	37083.000	0	0	2
2	2	3	164970.000	0	1	4
3	4	6	783045.000	0	0	4
4	2	4	1408274.000	2	3	5
5	2	5	2974774.000	4	0	0

Table 6: Cluster formation by cases for Katsina zone data using Complete Linkage method.

Number of clusters	Case										
	5		6		4		8		3		2
1	X	X	X	X	X	X	X	X	X	X	X
2	X		X	X	X	X	X	X	X	X	X
3	X		X	X	X		X	X	X	X	X
4	X		X		X		X	X	X	X	X
5	X		X		X		X	X	X		X

2 Tuberculosis, 3 Chickenpox, 4 Typhoid fever, 5 Malaria, 6 Cholera, 8 Measles

Table 7: Agglomeration Schedule for Katsina zone data using Centroid method.

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
1	3	8	37083.000	0	0	2
2	2	3	124805.750	0	1	3
3	2	4	699493.556	2	0	4
4	2	6	940245.625	3	0	5
5	2	5	2108153.400	4	0	0

By using the centroid method, the analysis result gives (Table 8) the following clusters: 1st Cluster: Malaria; 2nd Cluster: Malaria and Cholera 3rd Cluster: Malaria, Cholera and Typhoid Fever; 4th Cluster: Malaria, Cholera, Typhoid Fever and Chickenpox; 5th Cluster: Malaria, Cholera, Typhoid Fever, Chickenpox and Measles. Table 9 presents the agglomeration schedule that shows the stages involved in forming the clusters based on the Euclidean distances between the diseases.

By using ward method, the analysis result gives the following clusters: 1st Cluster: Malaria; 2nd Cluster: Malaria and Cholera; 3rd Cluster: Malaria, Cholera and Typhoid Fever; 4th Cluster: Malaria, Cholera, Typhoid Fever and Chickenpox; 5th

Cluster: Malaria, Cholera, Typhoid Fever, Chickenpox and Measles.

Table 11 presents the summary of cluster formation analysis for Katsina zone data and indicates that Malaria is more prevalent disease in Katsina zone irrespective of the cluster formation method employed. It is closely followed by Cholera using Single and Centroid methods and Typhoid fever using Complete and Ward methods. They ware also followed by Typhoid fever, Chickenpox and Measles using Single and Centroid methods and Cholera, Chickenpox and Measles.

Table 8: Cluster formation by cases for Katsina zone data using Centroid method.

Number of clusters	Case										
	5		6		4		8		3		2
1	X	X	X	X	X	X	X	X	X	X	X
2	X		X	X	X	X	X	X	X	X	X
3	X		X		X	X	X	X	X	X	X
4	X		X		X		X	X	X	X	X
5	X		X		X		X	X	X		X

2 Tuberculosis, 3 Chickenpox, 4 Typhoid fever, 5 Malaria, 6 Cholera, 8 Measles

Table 9: Agglomeration Schedule for Katsina zone data using Ward method.

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
1	3	8	18541.500	0	0	2
2	2	3	101745.333	0	1	4
3	4	6	493267.833	0	0	4
4	2	4	1378562.000	2	3	5
5	2	5	3135356.500	4	0	0

Cluster formation by cases for Katsina zone data using Ward method

Table 10: Cluster formation by cases for Katsina zone data using Ward method.

Number of clusters	Case										
	5		6		4		8		3		2
1	X	X	X	X	X	X	X	X	X	X	X
2	X		X	X	X	X	X	X	X	X	X
3	X		X	X	X		X	X	X	X	X
4	X		X		X		X	X	X	X	X
5	X		X		X		X	X	X		X

2 Tuberculosis, 3 Chickenpox, 4 Typhoid fever, 5 Malaria, 6 Cholera, 8 Measles

Table: 11: Summary of cluster formation analysis for Katsina zone data.

Clusters	**Methods for cluster formation**			
	Single Linkage	Complete Linkage	Centroid Method	Ward Method
1	5	5	5	5
2	5,6	5,4	5,6	5,4
3	5,6,4	5,4,6	5,6,4	5,4,6
4	5,6,4,3	5,4,6,3	5,6,4,3	5,4,6,3
5	5,6,4,3,8	5,4,6,3,8	5,6,4,3,8	5,4,6,3,8

2 Tuberculosis, 3 Chickenpox, 4 Typhoid fever, 5 Malaria, 6 Cholera, 8 Measles

Analysis and Interpretation Of Data From Daura Zone

The results in Table 12 Case Analysis of Katsina Zone data using Single Linkage Method are the

same by using the remaining methods: Complete Linkage Method, Centroid Method and Ward Method. In all the methods six diseases were observed and two diseases Leprosy and Tetanus were not considered because of their high Euclidean distance.

Table 12: Case Analysis of Daura Zone data using the Single Linkage Method.

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
6	75.0	2	25.0	8	100.0

Proximity Matrix Analysis

The proximity matrix analysis for Daura zone data using Single Linkage method, Complete Linkage Method, Centroid Method and Ward Method shows the relationship between any two of the diseases. The result in Table 13 is the same results for the remaining three methods. The results for all the methods show the distance between the diseases. Table 14 presents the agglomeration schedule that shows the stages involved in forming the clusters based on the Euclidean distances between the diseases. Using single linkage method, the analysis result (Table 15) gives the following clusters: 1st Cluster: Malaria; 2nd Cluster: Malaria and Cholera; 3rd Cluster: Malaria, Cholera and Typhoid Fever; 4th Cluster: Malaria, Cholera, Typhoid Fever and Measles; 5th Cluster: Malaria, Cholera, Typhoid Fever, measles and Chickenpox. Table 16 presents the agglomeration schedule that shows the stages involved in forming the clusters based on the Euclidean distances between the diseases. Cluster: Typhoid Fever; 2nd Cluster: Typhoid Fever and Malaria; 3rd Cluster: Typhoid Fever, Malaria and Cholera; 4th Cluster: Typhoid Fever, Malaria, Cholera, and Measles; 5th Cluster: Typhoid Fever, Malaria, Cholera, Measles and Chickenpox (Table 17).

Table 18 presents the agglomeration schedule that shows the stages involved in forming the clusters based on the Euclidean distances between the diseases. The results of Agglomeration Schedule for Daura zone data using Ward method was presented in Table 19 which gives the following clusters: 1st Cluster: Malaria; 2nd Cluster: Malaria and Cholera; 3rd Cluster: Malaria, Cholera and Typhoid Fever; 4th Cluster: Malaria, Cholera, Typhoid Fever and Measles; 5th Cluster: Malaria, Cholera, Typhoid Fever, measles and Chickenpox. Table 20 presents the agglomeration schedule that shows the stages involved in forming the clusters based on the Euclidean distances between the diseases.

The result of Cluster formation by cases for Daura zone data using Ward method was presented in Table 21. The analysis gives the following clusters: 1st Cluster: Typhoid Fever; 2nd Cluster: Typhoid Fever and Malaria; 3rd Cluster: Typhoid Fever, Malaria and Cholera; 4th Cluster: Typhoid Fever, Malaria, Cholera and Measles; 5th Cluster: Typhoid Fever, Malaria, Cholera, Measles and Chickenpox. Table 22 presents the summary of cluster formation analysis for Daura zone data and indicates that Malaria is more prevalent disease using Single and Centroid methods and Typhoid fever using Complete and Ward Methods. They are closely followed by Cholera using Single and Centroid methods and Malaria using Complete and Ward methods. They were also followed by Typhoid fever, Measles and Chickenpox for Single and Centroid methods and Cholera, Measles and Chickenpox for Complete and Ward methods.

Table 13: Proximity matrix of Daura zone data using Single Linkage method.

Case	Squared Euclidean Distance					
	2	3	4	5	6	8
2:TUBERCULOSIS	.000	92914.000	951258.000	4522662.000	3199791.000	601479.000
3:CHICKENPOX	92914.000	.000	1010698.000	3914086.000	3262047.000	246735.000
4:TYPHOIDFEVER	951258.000	1010698.000	.000	2942846.000	997701.000	1344831.000
5:MALARIA	4522662.000	3914086.000	2942846.000	.000	3015955.000	3029029.000
6:CHOLERA	3199791.000	3262047.000	997701.000	3015955.000	.000	3544414.000
8:MEASLES	601479.000	246735.000	1344831.000	3029029.000	3544414.000	.000

2 Tuberculosis, 3 Chickenpox, 4 Typhoid fever, 5 Malaria, 6 Cholera, 8 Measles

Table 14: Agglomeration Schedule for Daura zone data using Single Linkage method.

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
1	2	3	92914.000	0	0	2
2	2	8	246735.000	1	0	3
3	2	4	951258.000	2	0	4
4	2	6	997701.000	3	0	5
5	2	5	2942846.000	4	0	0

Table 15: Cluster formation by cases for Daura zone data using Single Linkage method.

Number of clusters	Case										
	5		6		4		8		3		2
1	X	X	X	X	X	X	X	X	X	X	X
2	X		X	X	X	X	X	X	X	X	X
3	X		X		X	X	X	X	X	X	X
4	X		X		X		X	X	X	X	X
5	X		X		X		X		X	X	X

2 Tuberculosis, 3 Chickenpox, 4 Typhoid fever, 5 Malaria, 6 Cholera, 8 Measles

Table 16: Agglomeration Schedule for Daura zone data using Complete Linkage method

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
1	2	3	92914.000	0	0	2
2	2	8	601479.000	1	0	5
3	4	6	997701.000	0	0	4
4	4	5	3015955.000	3	0	5
5	2	4	4522662.000	2	4	0

Table 17: Cluster formation by cases for Daura zone data using Complete Linkage method.

Number of clusters	Case										
	5		6		4		8		3		2
1	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X		X	X	X	X	X
3	X		X	X	X		X	X	X	X	X
4	X		X		X		X	X	X	X	X
5	X		X		X		X		X	X	X

2 Tuberculosis, 3 Chickenpox, 4 Typhoid fever, 5 Malaria, 6 Cholera, 8 Measles

The analysis result gives the following clusters: 1st

Table 18: Agglomeration Schedule for Daura zone data using Centroid method.

	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
1	2	3	92914.000	0	0	2
2	2	8	400878.500	1	0	3
3	2	4	997692.556	2	0	4
4	2	6	2485493.563	3	0	5
5	2	5	2874840.880	4	0	0

Table 19: Cluster formation by cases for Daura zone data using Centroid method.

Number of clusters	Case										
	5		6		4		8		3		2
1	X	X	X	X	X	X	X	X	X	X	X
2	X		X	X	X	X	X	X	X	X	X
3	X		X		X	X	X	X	X	X	X
4	X		X		X		X	X	X	X	X
5	X		X		X		X		X	X	X

2 Tuberculosis, 3 Chickenpox, 4 Typhoid fever, 5 Malaria, 6 Cholera, 8 Measles

Table 20: Agglomeration Schedule for Daura zone data using Ward method.

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
1	2	3	46457.000	0	0	2
2	2	8	313709.333	1	0	5
3	4	6	812559.833	0	0	4
4	4	5	2632543.333	3	0	5
5	2	4	5446074.333	2	4	0

Table 21: Cluster formation by cases for Daura zone data using Ward method.

Number of clusters	Case										
	5		6		4		8		3		2
1	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X		X	X	X	X	X
3	X		X	X	X		X	X	X	X	X
4	X		X		X		X	X	X	X	X
5	X		X		X		X		X	X	X

2 Tuberculosis, 3 Chickenpox, 4 Typhoid fever, 5 Malaria, 6 Cholera, 8 Measles

Table: 22: Summary of cluster formation analysis for Daura zone data.

Clusters	Methods for cluster formation			
	Single Linkage	Complete Linkage	Centroid Method	Ward Method
1	5	4	5	4
2	5,6	4,5	5,6	4,5
3	5,6,4	4,5,6	5,6,4	4,5,6
4	5,6,4,8	4,5,6,8	5,6,4,8	4,5,6,8
5	5,6,4,8,3	4,5,6,8,3	5,6,4,8,3	4,5,6,8,3

2 Tuberculosis, 3 Chickenpox, 4 Typhoid fever, 5 Malaria, 6 Cholera, 8 Measles

Comparism of the analysis and interpretation from the different zones

Table 23 presents the comparison of the analysis for the two different zones using single Linkage Method; the 1st Cluster is Malaria which is more prevalent in all the two zones. The 2nd Cluster is Malaria and Cholera and in the 5th Cluster we observed Measles in Katsina Zones while in Daura Zone is Chickenpox.

Table 24 presents the comparison of the analysis for the two different zones using Complete Linkage Method, the most prevalent disease is Malaria in Katsina Zone While in Daura Zone we observed Typhoid Fever. In the 2nd Cluster we observed Typhoid Fever and Malaria in Daura Zone and Malaria and Typhoid Fever in Katsina Zone. In the 5th Cluster we observed Measles in Katsina Zone while in Daura Zone we observed Chickenpox. Table 25 presents the

comparison of the analysis for the two different zones using Centroid Method; the 1st Cluster is Malaria which is more prevalent in all the two zones. The 2nd Cluster is Malaria and Cholera and in the 5th Cluster we observed Measles in Katsina Zone while in Daura Zone is Chickenpox. Table 26 presents the comparison of the analysis for the two different zones using Ward Method, the most prevalent disease is Malaria in Katsina zone while in Daura Zone is Typhoid Fever. The 2nd Cluster is Typhoid Fever and Malaria in Daura Zone and Malaria, Typhoid Fever in Katsina Zone. In the 5th Cluster is Measles in Katsina Zone while in Daura Zone is Chickenpox.

Table 23: Clusters established using the Single Linkage method for the two zones

Diseases by Prevalence	Daura Zone	Katsina Zone
1	Malaria	Malaria
2	Malaria, Cholera	Malaria, Cholera
3	Malaria, Cholera, Typhoid Fever,	Malaria, Cholera, Typhoid Fever
4	Malaria, Cholera, Typhoid Fever, Measles	Malaria, Cholera, Typhoid Fever, Chickenpox
5	Malaria, Cholera, Typhoid Fever, Measles, Chickenpox,	Malaria, Cholera, Typhoid Fever, Chickenpox, Measles

Table 24: Clusters established by Complete Linkage method for the two zones

Diseases by Prevalence	**Daura Zone**	**Katsina Zone**
1	Typhoid Fever	Malaria
2	Typhoid Fever, Malaria	Malaria, Typhoid Fever
3	Typhoid Fever, Malaria, Cholera,	Malaria, Typhoid Fever, Cholera
4	Typhoid Fever, Malaria, Cholera, Measles	Malaria, Typhoid Fever, Cholera, Chickenpox
5	Typhoid Fever, Malaria, Cholera, Measles, Chickenpox,	Malaria, Typhoid Fever, Cholera, Chickenpox, Measles

Table 25: Clusters established by Centroid method for the two zones

Diseases by Prevalence	**Daura Zone**	**Katsina Zone**
1	Malaria	Malaria
2	Malaria, Cholera	Malaria, Cholera
3	Malaria, Cholera, Typhoid Fever,	Malaria, Cholera, Typhoid Fever,
4	Malaria, Cholera, Typhoid Fever, Measles	Malaria, Cholera, Typhoid Fever, Chickenpox
5	Malaria, Cholera, Typhoid Fever, Measles, Chickenpox,	Malaria, Cholera, Typhoid Fever, Chickenpox, Measles

Table 26: Clusters established by the Ward method for the two zones.

Diseases by Prevalence	**Daura Zone**	**Katsina Zone**
1	Typhoid Fever	Malaria
2	Typhoid Fever, Malaria	Malaria, Typhoid Fever
3	Typhoid Fever, Malaria, Cholera,	Malaria, Typhoid Fever, Cholera
4	Typhoid Fever, Malaria, Cholera, Measles	Malaria, Typhoid Fever, Cholera Chickenpox
5	Typhoid Fever, Malaria, Cholera, Measles, Chickenpox,	Malaria, Typhoid Fever, Cholera Chickenpox, Measles

RESULTS FOR THE TEST OF INDEPENDENCE

Table 27: Test for Independence.

	Daura	Katsina	Total
Single Linkage	5 (5)	5 (5)	10
Complete Linkage	5 (5)	5 (5)	10
Centroid Method	5 (5)	5 (5)	10
Ward Method	5 (5)	5 (5)	10
	20	20	40

Table .27 presents the calculated chi-square as follows:

$\chi^2 = (5 ï 5)^2/5 + é é (5 ï 5)^2/5 = 0$

The tabulated chi-square is as follows:

$\chi^2_{\alpha(r-1)(c-1)} = \chi^2_{0.05,6} = 1.64$

Decision: we H_0 and conclude that the number of cluster formation tends to vary from one zone to another when different methods are employed.

CONCLUSION

The result of the cluster formation shows that Malaria is more prevalent in all of the two zones, followed by Cholera and Typhoid fever using the Single Linkage and Centroid methods. By using the Complete Linkage and Ward methods the results showed that Malaria is the most prevalent followed by Typhoid fever and Cholera in Katsina zone, while in Daura zone Typhoid fever is more prevalent followed by Malaria and Cholera. The results of the study show that Malaria is more prevalent in all the two zones, and by comparing the different methods employed, these is followed by Typhoid fever and Cholera. Although, some of the methods results differed slightly but the interesting result are that malaria, Typhoid fever and cholera are the most prevalent diseases in these two zones. We conclude from the chi-square test that the numbers of cluster formation tend to vary from one zone to another when different methods are employed. The prevalent of these diseases may be due to the nature of some areas in the state, where we have rivers, ponds where the causative agent (mosquitoes) can breed easily. Some of these diseases may also come from contaminated food, water and even fruits such as mangoes, cashew fruits, pawpaw etc. usually caused by housefly and tsetse fly.

RECOMMENDATIONS

Malaria being the most prevalent disease followed by Typhoid fever and Cholera in almost all the two zones in Katsina State, therefore there is the need for government or authority concerned to put in place sound programmes for the eradication of such diseases and provides welfare services such as drainage system, environmental protection and good pipe borne water.

REFERENCES

Blanchette, C. and Marks, J.R. (2000). Analysis of Gene Expression of Cancer Patients Discovery and Predictionò. *Gene Expression Monitoring Sci.* **286:** 531 - 537.

Bloom, A. (1963). *Toohey's Medicine for Nurses.* Whittington London.

Davidson, S. (2006). *Principle and Practice of Medicine, 20th Edition.* Edinburg, New York.

Everitt, B.S. (1974), *Cluster Analysis*. U.K.: Heinemann Educational Books Limited

FGN. (2007), Legal Notice on Publication of the 2006 Census Report. Federal Government of Nigeria official Gazette, 4(94): 1-8.

Gulumbe, S.U., Bakar, A.B. and Dikko, H.G. (2008). Classification of some HIV/AIDS Variables, a multivariate approach. *Res. J. Sci.* **15:** 24 ï 30.

Hardle, W. and Simar, L. (2007). *Applied Multivariate Statistical Analysis.* Spring Berlin Heidelberg, New York.

Hartigan, J.A. (1972). Direct clustering of a data matrix. *J. Am. Statis. Assoc.* **67:** 123- 129.

Mclaren, B. (1999). Probability Model based on bivariate histograms, measuring red blood cell volume and haemoglobin content. *7th International workshop on statistics.* Los Gatos C.A Morgan Kanfmann.

Rencher, A.C. (2002). *Method of Multivariate Analysis* 2nd edition, John Wiley & Sons Inc New York.

Solovyov, A., Palacious, G., Briese, T., Lipkin, W.I. and Rabadan, R. (2009). Cluster Analysis of the Origins of the New Influenza A (H1N1) Virus. *Eur. Surveillance J.* **14:** *38 – 41.*