International Journal of Mathematics and Statistics Studies
Vol.5, No.1, pp.29-44, February 2017
Published by European Centre for Research Training and Development UK (www.eajournals.org)

CLUSTER ANALYSIS OF THE INCIDENCES ON HIV IN NIGERIA

Dr. Yahaya H.U
Department of Statistics
University of Abuja, Abuja Nigeria

Mr. Raheem Kola
Department of Statistics
University of Abuja, Abuja Nigeria

ABSTRACT: *Data clustering is a vital tool when it comes to understanding data items with similar characteristics in a data set for the sake of grouping. Clustering may be for understanding or utility. Clustering for understanding, which is the focus of this work deals with grouping items with common characteristics in order to better understand a dataset and to identify possible or pre-interest sub-groups that could be formed from such data. The HIV prevalence statistics in Nigeria is measured bi-annually across 36 states and FCT which were zoned under 6 geo-political zones happens to be a suitable data to implement this subject matter. Cluster Analysis was implemented through the general methods of Hierarchical (agglomerative nesting) and Partitioning methods (K-Means). These techniques where implemented on the platform of R (Statistical Computing Language) to cluster HIV prevalence rate in Nigeria so as to find out states that could be considered same category and to investigate the concentration of the disease in respect to geo-political zones. Relative type of validation was used for cluster validation (a mechanism for evaluating the correctness of clustering).*

KEYWORDS: Data, Clustering analysis, HIV, Pregnant women, Nigeria,

INTRODUCTION

Cluster analysis divides data into groups called clusters such that items in the same cluster are similar while items in different cluster are distinct, the essence of such grouping may be for *understanding* or for *utility*. The first case, clustering for understanding has to do with understanding the behaviour of data or in some respect knowing the characteristics that data items may have in common, this is the focal point of this work. Meanwhile, clustering for utility has to do with classification of data for further usage or analysis.

Over the years, the world had suffered several deaths due to a virus known as HIV, several institutes, governments and international organisations had invested a lot in providing measures to curb the spread of this disease, but till date HIV still posed a significant challenge among other diseases particularly in Africa. Nigeria as a case study has second highest number of people living with HIV (*CIA World Factbook, 2012*) with an estimate of 3.3 million of its populace which amounts to about 2% of the entire population. The country has 36 states zoned under 6 geo-political zones. The HIV prevalence rate is measured periodically across these states by means of a sentinel survey.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

This research work applied the technique of cluster analysis on the platform of R (statistical computing) to classify the spread of this virus among the populace of Nigeria. Classification of this information would help to take better decisions as to what part of the country needs certain attention.

HIV Prevalence in Nigeria

The HIV and AIDS pandemic have constituted the greatest health challenge in Nigeria over time. In 2012 Nigeria was adjudged to have the second highest burden of HIV in the world after South Africa. Since the first reported case of HIV and AIDS in Nigeria in 1986, the epidemic has continued to unleash a huge blow on the commonwealth with about 3.3 million Nigerians currently infected. To respond to this epidemic, the Federal Government of Nigeria put in place several programmes aimed at controlling and mitigating its effect. The sole purpose of these intervention programmes is usually the continuous monitoring of the HIV epidemic via a biennial sentinel survey among pregnant women at finishing antenatal clinics in Nigeria.

In the African region, active HIV sero-surveillance using pregnant women attending ante-natal clinics as the survey population is employed in line with the World Health Organization (WHO) plus the Joint United international locations Programme on HIV and AIDS (UNAIDS). The HIV sentinel sero-surveillance survey has been conducted biennially in Nigeria since 1991.The HIV prevalence in Nigeria had been on a consistent increase from 1.8% in 1991 to 5.8% in 2001 before a decline to 5% in 2003 and 4% in 2005. The report of the 2005 survey further reaffirms that no state or community is spared this epidemic. There are wide variations in HIV prevalence among states and between urban and rural places across the commonwealth. Data resulting from sentinel survey could be inconclusive to get ready direct comparisons between aggregate figures acquired inside a several surveys due to differences in location and counts of survey sites.

METHODOLOGY

Cluster analysis is a study of dividing data items into groups such that the elements of each group are homogenous as much as possible. There are basically two major approaches used for grouping in cluster analysis, these are *Hierarchical* and *Partitioning method*.
For the sake of this work, Agglomerative nesting and K-Means approach which are methods of Hierarchical and Partitioning clustering would be our focus

Agglomerative Nesting (Hierarchical)

Agglomerative nesting often called AGNES uses an algorithm in a bottom-up manner till a single rooted tree like diagram (dendogram) is formed. The algorithm is as follows:

1. Initially, put each article in its own cluster.
2. Among all current clusters, pick the two clusters with the smallest distance.
3. Replace these two clusters with a new cluster, formed by merging the two original ones.
4. Repeat the above two steps until there is only one remaining cluster in the pool.

Distance Measure

In order to compute clustering, a measure of dissimilarity between sets of observations is required. This is sometime refers to as proximity matrix (a matrix of distance measure which could either be similarity measure or dissimilarity measure as in the case of hierarchical clustering). The proximity matrix is achieved by use of an appropriate metric such as Euclidean distance and Manhattan distance measures.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

Euclidean distance:
Euclidean distance is an inter point distance, it takes the magnitude of the expression data into account and therefore, preserves more information about the data. Euclidean distance is the length of the shortest path between two points and required standardization when used. Mathematically given as:

$$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2} \quad - \quad - \quad - \quad 3.1$$

Manhattan Distance
The Manhattan distance also known as the City-Block distance is the sum of distances along each dimension. This distance measure corresponds to the distance of travel between two points. By formula:

$$\|a-b\|_1 = \sum_i |a_i - b_i| \quad - \quad - \quad - \quad - \quad 3.2$$

The cluster analysis computation of this work uses the Manhattan distance.

Ward's Linkage Criterion
Linkage criterion determines how clusters should be formed; it computes the distance between clusters.Ward's method is a criterion applied in hierarchical cluster analysis. Ward's minimum variance method is a special case of the objective function approach originally presented by Joe H. Ward, Jr. (1963). Ward suggested a general agglomerative hierarchical clustering procedure, where the criterion for choosing the pair of clusters to merge at each step is based on the optimal value of an objective function
Ward's minimum variance criterion minimizes the total within-cluster variance. At each step the pair of clusters with minimum cluster distance are merged

$$d_{ij} = d(\{X_i\}, \{X_j\}) = \|X_i - X_j\|^2. \quad - \quad - \quad 3.3$$

This method is distinct from other methods because it uses an analysis of variance approach to evaluate the distances between clusters. In general, this method is considered very efficient.

Agglomerative Coefficient
Agnes computes a coefficient, called Agglomerative Coefficient (AC), which measures the clustering structure of the data set.
Agglomerative Coefficient is a dimensionless quantity, varying between 0 and 1. When AC is close to 1 there is an indication that a very clear structuring has been found. Otherwise when AC is close to 0 it indicates that the algorithm has not found a natural structure. In other words, the data consists of only one big cluster.

K-Means (Partitioning)
This is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. K-means is a prototyped based approach that requires that the number of cluster be specified before the analysis is conducted.
Determining number of cluster to specify is a major challenge of using the k-means. There are several theories on how to compute and obtain the appropriate number; these would not be discussed since it is beyond the scope of this work.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

There is a Rule of thumb for convenience sake which is an estimate given as:

$k \approx \sqrt{n/2}$ - - - - 3.4

Where: k – Number of clusters to specify

n – Sample size

Agglomerative Nesting (Before Standardization):

The following R code computes the Agglomerative Hierarchical Clustering using the default (distance measure = average and variables are not standardized) parameters as contained in the cluster library in R. At this point the data was not standardized.

Explanation of Code Listing 4.0 A:

The first line of code is used to invoke the cluster library (package that contains code for computing cluster analysis).

```
library(cluster)
```

The next line reads the data shown above (table 4.0) from text file named HIV.txt into an array variable called data. Header=True means that the columns should be named and row.names=1 implies that the rows should also have title (state names).

The line of code below creates an object called data.agnes (an instance of agnes class). Where agnes implies Agglomerative nesting

```
data.agnes <- agnes(data)
```

The last line (data.agnes) executes the agnes computation.

RESULT/FINDINGS

Below is the output (result) of the execution of the code listing 4.0. The execution produced a summary Table 1 as shown below

```
Call:    agnes(x = data)
Agglomerative coefficient:  0.8885348
Order of objects:
 [1] Abia        Borno       Sokoto      Delta       Imo
 [6] Lagos       Bauchi      Kano        Ondo        Oyo
[11] Katsina     Yobe        Kebbi       Kwara       Zamfara
[16] Osun        Ogun        Ekiti       Jigawa      Adamawa
[21] Edo         Kogi        Taraba      Niger       Anambra
[26] Rivers      Bayelsa     Enugu       Ebonyi      Gombe
[31] Plateau     Kaduna      Akwa Ibom   Cross River Nasarawa
[36] FCT         Benue
Height (summary):
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0700  0.4742  0.8066  1.2940  1.4820  7.2070
```

Source: R console Output

Table 1: Interpretation of R Output:

Agglomerative Coefficient (AC):

Table 1 summarizes the result of the computed analysis. The Agglomerative coefficient measures the clarity of the structure of clusters formed as shown in Table 1. Agglomerative coefficient of the computed agnes was 0.8885348 (i.e. about 89%) indicates a clear structure or a structure

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

close to a natural structure since the value is close to unity (see section 3.03 above). The AC of 89% therefore means that the clusters generation formed a very good and clear structure.

a. Order of Objects:

This shows the order in which the objects (States) were selected to form the clusters, that is the order in which the agglomerative algorithm selected the States at the process of forming the cluster. It is noticed that States with a reasonably low or average prevalence rate were found at the top of the ordering while States with high prevalence were towards at the end. Though this seem to look as if the States were arranged in ascending order of prevalence rate, but it does not exactly appear that way. Further result shed more light to the ordering. This ordering appears as nodes of the dendogram in Table 1

b. Height Summary:

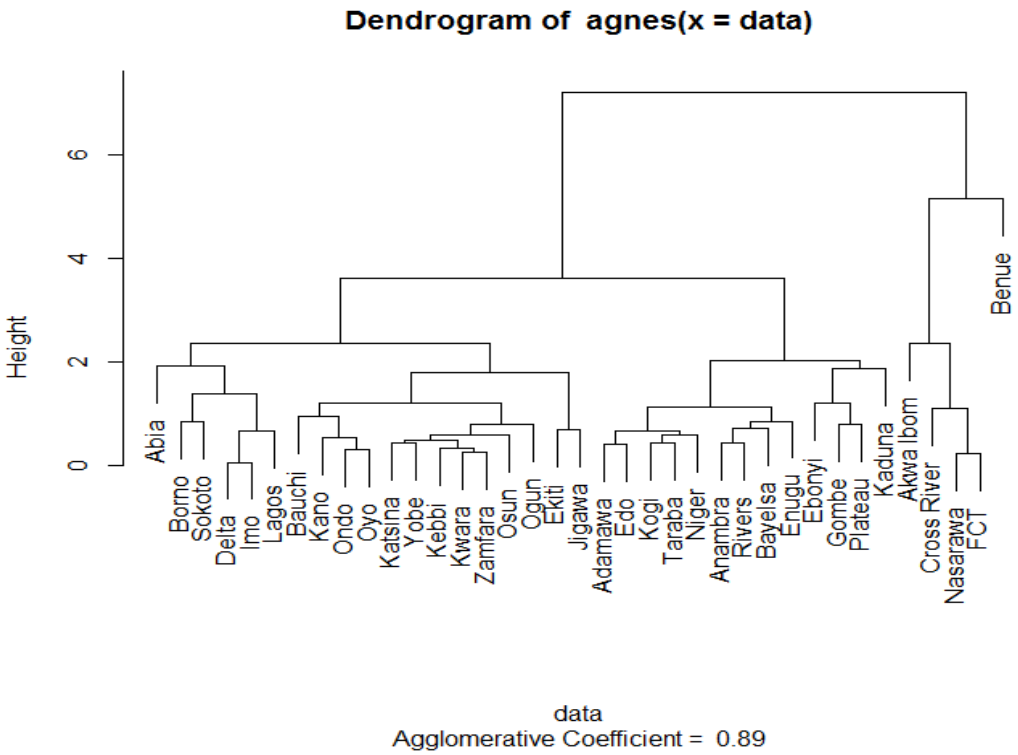
The height measures the distance between clusters formed, this forms the y-axis of the hierarchical structure below. The minimum distance within the clusters formed is as low as 0.0700 and a maximum as high as 7.2070. This implies that the nearness or farthest of any two cluster is within the range of the stated values. Objects (States) in the same clusters are naturally expected to have minimal height difference. Clusters with a low distance would likely have more members with close level of prevalence than Clusters with a distance far apart.
The average (mean) 1.294 is the average distance between clusters

In order to see the graphs of the analysis, the data.agnes object is passed into a function called *plot( )* as in **Code listing 4.0 B**. The execution produces a dendogram as shown in Table 1

Dendogram of agnes

Dendogram was used to depict the result of hierarchical clustering by grouping similar items in the same cluster in a tree-like manner. As we move upward the dendogram the level of Homogeneity within cluster members increases and decreases by downward movement. Items in the same cluster were grouped due to common characteristics in their prevalence rate.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)



Source: R Console Output
Figure 1: Dendogram of AGNES

Interpretation of R Output:
The x-axis consists of the names of States according to the order in which they were selected to form clusters. Choosing the number of clusters is subjective since number of clusters depends on a given height (x-axis). The number of clusters varies as we move along the height upwardly or downwardly.

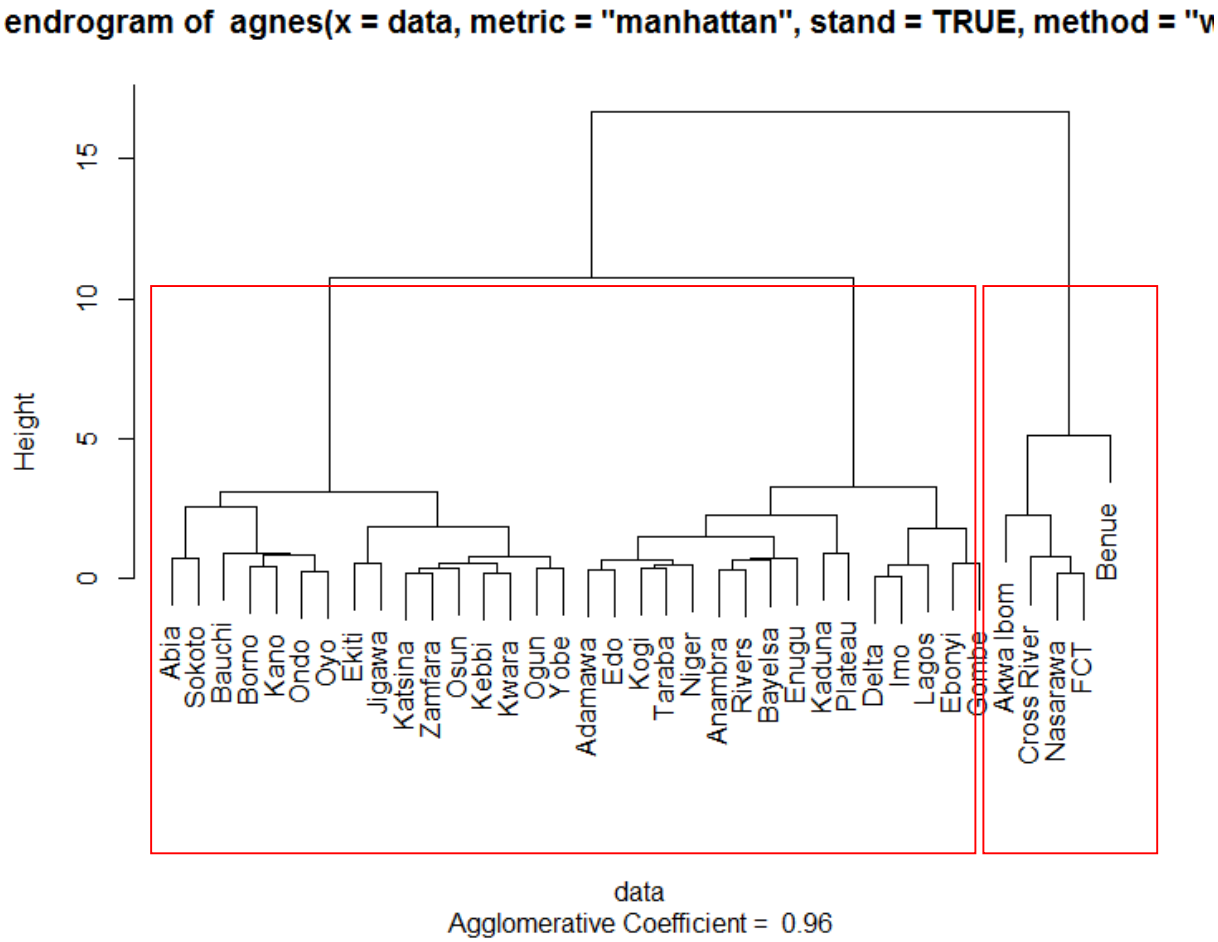
The height value labeled 0 to about 8 on the y-axis, recall that we have a minimum distance of 0.0700 and a maximum of 7.2070 as stated in summary table 1 above. The optimal number of clusters would be determined based on a criterion to maximize distance between clusters and yet identify distinct group. Interpreting the clusters could be sometimes difficult. A better result was obtained when the data was standardized (See table 2 **R Output 4.1** below) for further interpretation.

Agglomerative Nesting (after Standardization)
After computing the agglomerative nesting (agnes) with the un-standardized data, the analysis was re-computed with the standardized data this is to ensure that one variable do not dominate the analysis.The code listing 4.1 was used to execute the agnes specifying the distance measure (Manhattan) and enforcing that the data be standardized with the Ward's method of linkage as in the next line.

Interpretation of R Output 4.0 A:
The summary table (**R Output 4.1 A)** above shows the summary of agnes after standardizing the variables. After standardizing the data, the Agglomerative Coefficient (AC) increased from 89% to 96% meaning that we have a better structure than the first one. The height range in the earlier

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

result was 0.0700 to 7.2070, but we now have a range of about 0.04004 to 16.7100 which implies a wider difference between clusters. This enhances the aim of maximizing the inter-cluster distance. Several changes were also noticed in the order of objects.

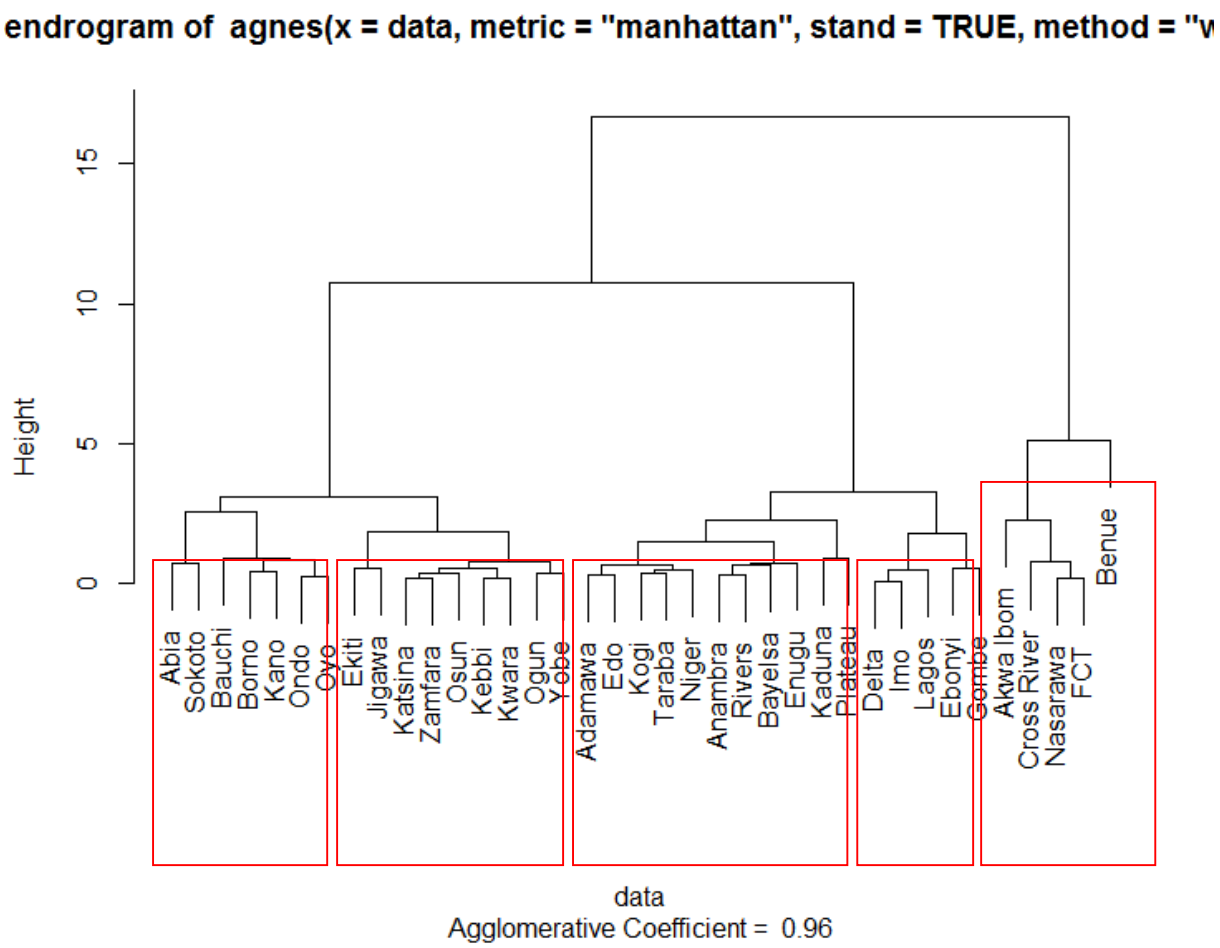


Source: R Console Output
Figure 2: Two Cluster Solution

Interpretation of R Output 4.1 C:
The height on the y-axis indicates the distance between one cluster and the other. Looking at this structure from the top, we would see that we generally have two major groups from the dendogram. This is evident at the height of 12, if we move from a height of 12 on the y-axis across the x-axis we would cross two lines, which indicate the two major groups that we have in the result. We call this a two cluster (2-cluster) solution. This implies that Akwa-Ibom, Cross-River, Nasarawa, F.C.T and Benue are in the same category against the other 33 States as marked with the red rectangles. A reflection of this is showed on the average prevalence rate; these states have the highest prevalence rates.

As we move down the dendogram and read the line in the same manner used to obtain a 2-Cluster solution above, we would encounter more clusters. The dendogram was further analysed below.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)



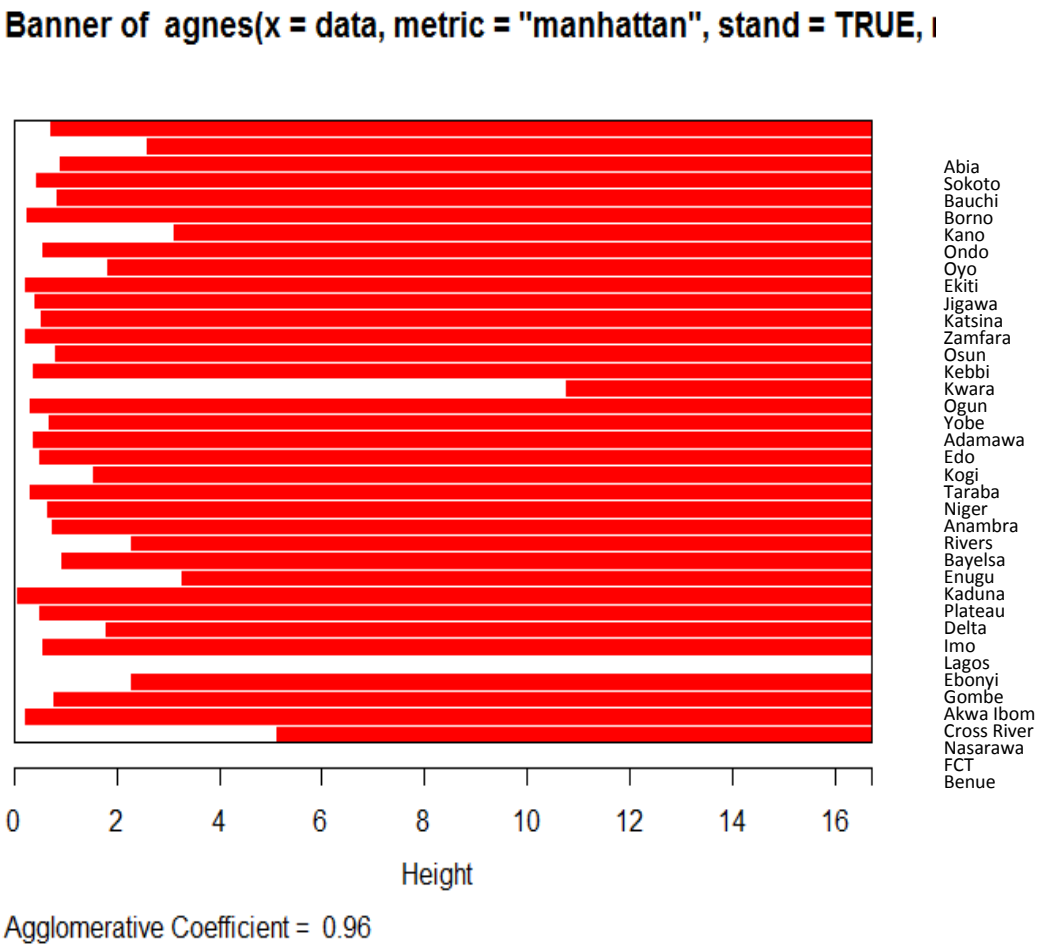
Source: R console Output
Figure 3: Five Cluster Solution

Interpretation of R Output 4.1 D:
Furthermore, by taken a closer look at the two-cluster solutions, the cluster on the left is further divided into two sub-groups at about a height of 10 as we moved downward from 12, at this level we would say we have three-cluster solutions. Further look into the dendogram reveals that we have a five cluster solution as shown in the figure **R Output 4.1 D** above. At this point, the States are distributes in 5 clusters of sizes 7, 9, 11, 5, and 5 taking each group from the left.

Banner Plot

The banner plot is an alternative to dendogram. The white area on the left of the banner plot (**R Output 4.0 b**) represents the un-clustered data while the white lines that stick into the red shows the heights at which the clusters were formed. We considered the dendogram preferable though interpreting the banner plot would lead to same conclusion. The choice using the banner plot is determined by the analyst, we found the dendogram to be better for this reason we stick to it.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)



Source: **R Output 4.1 C**
Figure 4: Banner Plot

K-Means (partitioning)
K-means is a prototype based technique; the number of clusters to be formed needs to be specified before the computation as discussed earlier. By applying the Rule of thumb, with sample size of 37, our k (number of clusters) is around 4.30. Since choosing a number of clusters in k-means could be decided as willed, k=5 was used. The following code listing was used to scale the average HIV prevalence data, the output of the scaled data appears in R output 4.2 A

Code listing 4.2 A

K MEANS
The output below shows the scaled data after been scaled. Scaling the variables ensures that the variables are on the same metric.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

```
K-means clustering with 5 clusters of sizes 13, 5, 7, 10, 2

Cluster means:
  X1999_2003 X2005_2010
1  0.3009821  0.3436609
2  1.9106795  1.8553624
3 -0.3790514 -0.2508853
4 -0.8173938 -0.9216925
5 -1.3194336 -1.3856407

Within cluster sum of squares by cluster:
[1] 3.04525944 4.79277415 1.41354797 0.58009386 0.04531497
 (between_SS / total_SS =  86.3 %)


Clustering vector:
       Abia     Adamawa   Akwa Ibom     Anambra      Bauchi     Bayelsa
          3           1           2           1           3           1
     Benue       Borno Cross River       Delta      Ebonyi         Edo
          2           3           2           3           1           1
     Ekiti       Enugu       Gombe         Imo      Jigawa      Kaduna
          5           1           1           3           5           1
      Kano     Katsina       Kebbi        Kogi       Kwara       Lagos
          4           4           4           1           4           3
  Nasarawa       Niger        Ogun       Ondo        Osun        Oyo
          2           1           4           4           4           4
   Plateau      Rivers      Sokoto      Taraba        Yobe     Zamfara
          1           1           3           1           4           4
       FCT
          2
```

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Adamawa	Akwa Ibom	Abia	Kano	Ekiti
Anambra	Benue	Bauchi	Katsina	Jigawa
Bayelsa	Cross River	Borno	Kebb	
Ebonyi	Nasarawa	Delta	Kwara	
Edo	FCT	Imo	Ogun	
Enugu		Lagos	Ondo	
Gombe		Sokoto	Osun	
Kaduna			Oyo	
Kogi			Yobe	
Niger			Zamfara	
Rivers				
Taraba				

Source: Author's Computation
Table 3: Cluster Membership for K-means Clusters

The output in the summary table above indicates 86.3% total variation, this implies that about 86% of the variation present in our data was explained by the clustering result. This shows that we have a reasonably good clustering as the clusters formed were able to account for 86% of the differences in the HIV prevalence rate.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

Cluster means indicates the centres of the clusters formed. It is the mean of the specific variables combined to form a particular cluster. K-means attempt to minimise variation (difference in prevalence rate) within members of a cluster and maximise it between clusters so that members in the same cluster are homogeneous as much as possible and heterogeneous to other clusters.
For clarity of what the centres shows we present the table 4 below:

Cluster Size	**X1999_2003**	**X2005_2010**	**Final Mean**
13	0.3009821	0.3436609	0.644643
5	1.9106795	1.8553624	3.7660419*
7	-0.3790514	-0.2508853	-0.6299367
10	-0.8173938	-0.9216925	-1.7390863
2	-1.3194336	-1.3856407	-2.7050743

Source: R console Output
Table 4: Cluster Means (centres) for K-means

The mean of the second cluster (3.766) in the above table is the most distinct among others, we observed that members (states) of this cluster have extreme prevalence rate as evident in our original data. It is also noticed that these states are exactly the same as found in the last cluster of our 5-Cluster solution, far right of the dendogram in **R Output 4.1 D** above.

The result of the K-means to a large extent agreed with the agglomerative nesting at k=5 no of cluster. Clusters formed by both analysis (Agglomerative nesting and K-means) have much similarities, this is clear by comparing the Clustering Vector (**Table 3.**) and the five cluster solution in the Dendogram (**R Output 4.1 D**).

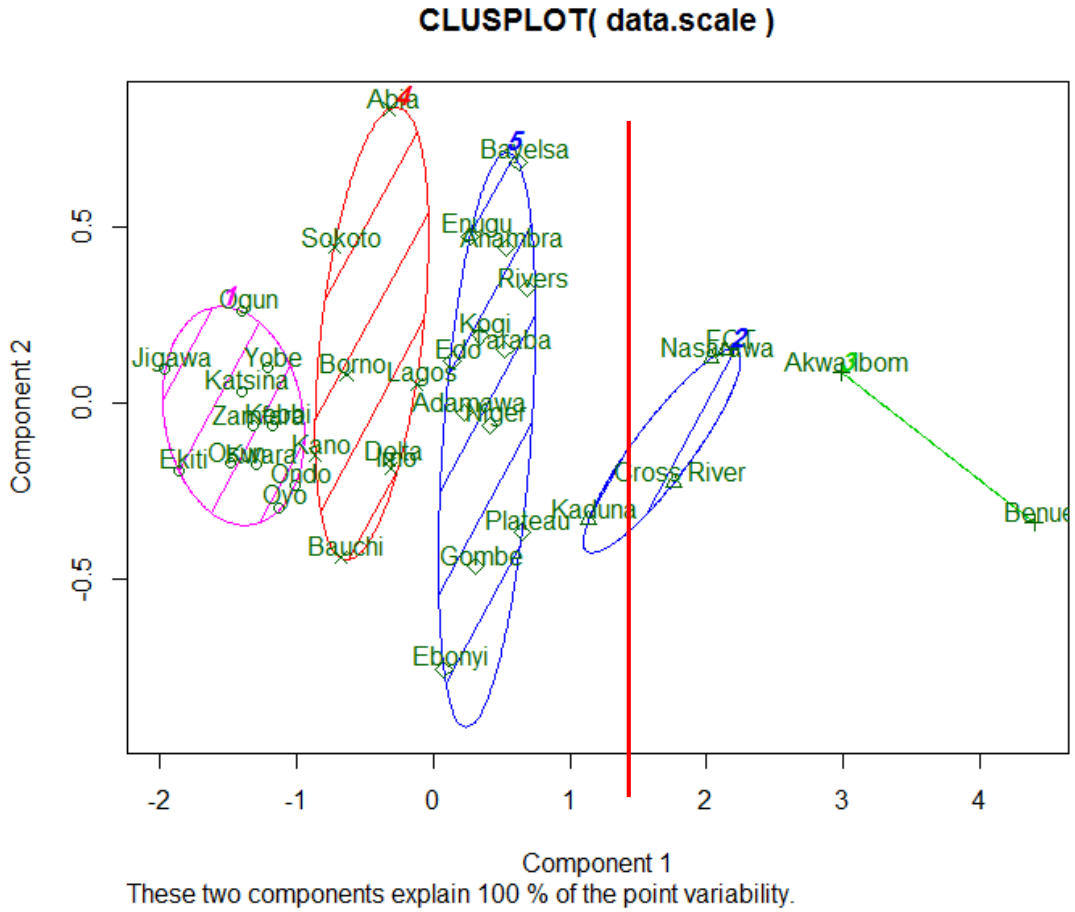
Clusplot Graph Function
The clusplot( ) is a function that display Cluster on a bi-component plot. It uses the PCA (Principal Component Analysis) approach on the given data. Principal component analysis (PCA) is a dimension reduction technique; it summarizes the information of all variables into a couple of new variables called *Components*. Each component is responsible for explaining certain percentage of the total variability. The clusplot uses the first two components which usually explain the maximum variation in the given data to display a principal plane.

The bi-plot obtained from clusplot is used to identify the effectiveness of clustering. Clearly separated cluster in the principal plane indicates successful clustering. On the other hand, a merged cluster implies unsuccessful clustering
The clusplot function takes the parameter of a k-means object and specification of colours to differentiate groupings or cluster membership. The code listing below was used to implement clusplot.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

Code listing 4.2 C

```
> fit <- kmeans(data.scale, 5)
> clusplot(data.scale, fit$cluster, color = TRUE, shade = TRUE, label=2,
lines=0)
```



Source: R console Output
Figure 5: Clustering Plot

Interpretation of R Output 4.2 A:
It is obvious from the plot that the objects (States) formed a kind of clusters on the principal plane, though this clusters differs from that of k-means, it is not a problem since the clustering plot aims to show the spread of the objects on the plane, we are interested in knowing if these objects worth having a substantial grouping as claimed by the K-means by looking at their spread on the plane.

The different colour shadings indicates different groups and objects that are likely to have same cluster membership, it is evident from this figure that the objects have sufficient groupings. The two components explained 100% of the total variability, meanwhile the first component explains much of the variation, this is customary to principal component analysis. The values on the axis are not necessarily important.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

An important observation is that states with high prevalence rate falls to the right of the red line at the centre of the clusplot. In fact, the farther the state to the middle line the higher the prevalence rate. Thus, the plot clearly shows States that policies makers should give more attention.

Cluster Validation

The result of our clustering was subjected to validation by means of Relative approach as discussed earlier (Section 2.3) K-means produced clusters of sizes 13, 5, 7, 10, and 2 while the Hierarchical produced clusters of sizes 7, 9, 11, 5, and 5. If we compare members of a given cluster in k-means to the members of a similar cluster in hierarchical we would see that there are lots of similarity between the clusters produced by the two methods. For clarity, we assign a group number to each cluster in both methods as follows:

Group No.	Group 1	Group 2	Group 3	Group 4	Group 5
Cluster Size	13	5	7	10	2

Source: Author's Computation

Table 5: K-means Clusters:

Group No.	Group 1	Group 2	Group 3	Group 4	Group 5
Cluster Size	11	5	7	5	5

Source: Author's Computation

Table 6: Agglomerative Nesting Clusters:

The table below summarizes the relationship between the clusters generated by K-means and the Agglomerative Nesting.

Clusters	K Means (Cluster size)	Hierarchical (Cluster size)	Difference in Cluster Size	No. of Agreed Members	No. of Different Members
Group 1	13	11	2	10	3
Group 2	5	5	0	5 (All)	0
Group 3	7	7	0	4	3
Group 4	10	9	1	6	4
Group 5	2	5	3	0	7

Source: R console Output

Table 7: Comparison Table for K-Means and Hierarchical Clusters Membership

Starting with Group 1, K-Means and Agnes have 13 and 11 cluster sizes respectively, whereas 10 of the members are equal in both clusters (that is 10 of the States in Group 1 cluster of K-Means is found in Group 1 cluster of Agnes). These include Adamawa, Anambra, Bayelsa, Edo, Enugu, Kaduna, Kogi, Niger, Rivers, and Taraba States. But Ebonyi and Gombe were found in Group 1 of K-Means and Plateau in Group 1 of Agnes which implies a difference of 3 members, 2 from K-Means and 1 from Agnes.

There is a high level of similarity between the cluster sizes produced by the two methods as shown in the column "Difference in Cluster Size". In fact, they both have equal cluster sizes on two occasions (Group 2 and 3). Group 2 and 3 have equal number of members in both K-Means

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

and Agglomerative methods, though there is member difference in Group 3. A special case is the Group 2, they do not only have equal number of members, but they also have the same States (Akwa-Ibom, Cross-River, Nasarawa, FCT, and Benue) as cluster members.

Group 5 appears to be the only case where there no similar member in both cases. Therefore, both clusters have 7 distinct members in which 2 belong to K-Means and 5 to Agnes. Exact match is not expected as both methods uses different algorithm to make their clustering.
It is reasonable to say that the difference in cluster sizes in both K-means and Agnes is low and there exist a high level of similarity in clustering and cluster membership yielded by both methods, this indicates goodness in the clustering results obtained.

SUMMARY, RECOMMENDATION AND CONCLUSION

Summary

The HIV prevalence rate in Nigeria is measured bi-annually, pregnant women under antenatal care in survey sites are usually the major source of measuring this statistics. The HIV prevalence data used for this work covers the periods of 1999 to 2010, though this period was split into two variables, the first variable being the period between 1999 – 2003 and the second variable is between 2005 – 2010, each variable consists 3 years data point since the measurement is done bi-annually. The split was necessary to correct the possible anomalies that could be posed by inconsistent lag in the second variable (2005 – 2010), which also led to the use of average prevalence rate in our analysis (see Appendix I).

We intend to find possible classification for which we can determine States with common situation of the HIV spread. We conducted a Cluster analysis using the clustering schemes from two clustering methods, Hierarchical and the Partitioning methods.

Agglomerative nesting an approach of Hierarchical clustering, divides data items into homogeneous group by means of a proximity matrix and a linkage criterion, this agglomerative techniques was first used on the prevalence data. This resulted in a five cluster solution after detail study of the output and a hierarchical structure (dendogram) as shown in **R Output 4.1 D**. The 36 States were grouped into clusters of sizes 7, 9, 11, 5, and 5. This result was accepted to a large extent since it generated the number of clusters that coincides with the recent literatures by (*FMOH 2010*) as recommended by WHO.

In the case of Partitioning method, we used the K-means, a prototype based approach in which a priori information about number of cluster must be specified. In conducting the K-means clustering, we specified k (number of cluster) to be formed as five (k=5), the same number of clusters generated in the agglomerative approach which is also the approximated value for using the rule of thumb for determining the appropriate number cluster for K-means clustering. This was done to see how well we can rely on the grouping done by the Hierarchical procedure, another reason for this is for cluster validation. The result of the K-means agreed with the agglomerative to a large extent, since they both have more similarities than differences.

The Relative type of Cluster validation (a check mechanism) was used to critically compare the clustering produced by the two clustering shames to ascertain that we did not cluster in noise (that is clustering when it is not necessary). A comparison table Fig 4.0 was established in order to ease the task of validating the clustering result.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

Recommendation

The quality of data plays a vital role in any statistical analysis. It is highly important that the sentinel survey for acquiring the HIV prevalence rate be made consistent. There should be consideration for having more sources of acquiring the prevalence rate rather than patients under antenatal care only.

Conducting a cluster analysis is not an easy task, the fact that there are multiple ways of implementing cluster analysis in which combination of steps and possible alternatives depends on the choice of analyst makes it subjective. Meanwhile, each combination of steps and alternatives has its own weakness. It is important to use different clustering schemes on the data of interest in order to see the direction of the obtained result and to avoid misleading clustering. A tabular mechanism for comparing result of two clustering scheme used in Table 4.2 above provides an easy alternative for comparing clustering results from different clustering scheme. Further research works could develop numerical computations to enhance or optimize this approach and make the process faster.

A major aim of Governments and Agencies concerned with HIV issues is to minimize the spread of this disease as much as possible. The plot above (**R Output 4.2 C**) could be an easy way to clearly identify States that require more attention. More of the programmes such as awareness and education on HIV should be channelled to such States. Another point of interest that government and agencies in control of HIV in Nigeria should look into is the connection between bilateral relationships and prevalence rates; the clustering map (figure 4.4) gives an insight on this.

CONCLUSION

From the results obtained it can be said that we have a successful clustering that classified the spread of HIV among Nigerian states where North-Central and the South-South geopolitical zones are important areas that require more attention. Clustering Plot (R Output 4.2 C), a graph of K-means clustering reveals an important measure for identifying areas (states) under risk; this graph requires a closer look.

Despite the fact that K-means and Agglomerative clustering are different clustering techniques and choice of using any may be based on the type of data; the final results of these techniques for the same data shows lots of similarities.

REFERENCE

Alka A. (2011) Cluster Analysis Using R, "winter school of data mining techniques and tools for knowledge discovery in agricultural datasets"

Brian S. E., Morven L. Sabine L. and Daniel St. (2011). Cluster Analysis – 5th edition, John Wiley & Sons, Ltd. 2011, 61, 80-81

Jeromy A. (2007) Cluster Analysis & Factor Analysis, Lecture Note: 325-711

Kaufman, L. Reusseeuw P. (1990). Finding Groups in Data: An introduction to cluster analysis John Wiley and son, London. ISBN: 0471878766

Maindonald J. H. (2008) using R for data analysis and graphics- introduction, code and commentary

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

Marija N., IBM SPSS Statistics Procedure SPC pp361-363 http://www.norusis.com/pdf/spc v13. Pdf

Pang-Ning T. Michael S. and Vipin K. (2006) Introduction to data mining, Addison Wesley, Isbn-13:9780321321367.

Raza A. 425, Usman G. (462) and Aasim Saeed (464): data Clustering and its application, http:// members.tripod.com/asim saeed/paper.htmResearch Methods.

Steven M.H (2006) cluster analysis, department of Geology, University of georgia

The R development core team (2010) R: A language and environment for statistical computing, reference index version 2.11.1

Volker H and Sebastian W. (2000), Clustering techniques: A brief survey, institufur biomathematik and biometrie, 7.

Ward, J. H., Jr. (1963), Hierachical Grouping an objective function, journal of the American statistics association, 58, 236-244.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)