MNK Publication

International Journal of Latest Research in Science and Technology **ISSN (Online):2278-5299**
Volume2,Issue 1 :Page No.437-441 ,January-February (2013)
http://www.mnkjournals.com/ijlrst.htm

A HIERARCHICAL CLUSTER ANALYSIS AND SIMULATION OF STATE CAPITALS IN NIGERIA FOR TOURISM EXPLORATION

[1]Audu Isah, PhD , [2]Usman Abdullahi, [3]Victor O. Waziri PhD,
[1]Department of Mathematics/Statistics , Federal University of Technology, Minna-Nigeria.
[2]Academic Planning Unit.Federal University of Technology, Minna-Nigeria
[3]Department of Cyber Security Science; School of ICT, Federal University of Technology, Minna-Nigeria

Abstract: In this paper we conducted hierarchical cluster analysis with multiscale bootstrap resampling technique using average linkage method and correlation based dissimilarity matrix to classify state capitals according to their distances by Road from Federal capital, Abuja to the respective state capitals. The R package pvclust was used to assess the uncertainty in hierarchical clustering results. This was then applied to the standard normal approximation theory on the clusters whose standard errors were high or larger than 0.030. The cluster formations have an apparent error rate of 0.027.

Keywords: Hierarchical cluster analysis, average linkage, highlighted clusters, approximately unbiased (AU) p-values, dendrograms, multiscale Bootstrap.

1. Introduction

Cluster analysis is a powerful exploratory technique for discovering groups of similar observations within a data set. The idea of cluster analysis is to use values of variables to devise a scheme for grouping objects in such a way that similar objects will belong to the same group (in some sense or another) to each other than to those in other groups. Cluster analysis may be worthwhile in finding the true groups that are assumed to truly exist and may also be useful for data reduction. It is used in many fields of study, including machine learning, image analysis, pattern recognition, information retrieval and bioinformatics.

Cluster analysis can be achieved through many algorithms that differ significantly in their notion of what constitute a cluster and how to efficiently find them. Popular notions of clusters include groups with low distances among cluster members, dense areas of the data space, intervals or particular statistical distributions.

There are many typical cluster analysis models in use today. In this paper, emphasis is on the connectivity models. These are hierarchical clustering build models based on distance connectivity. This method is based on the core idea of objects being more related to nearby objects than to objects farther away. As such, these algorithms connect objects to form clusters based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram which explains where the common name "hierarchical clustering" comes from and provide an extensive hierarchy of clusters that merge with each other at certain distances (en.wikipedia.org/wiki/cluster_analysis).

Agglomerative hierarchic methods start at the leaves and successively merge clusters together. The process of merging clusters is based on the estimation of observed similarity/ dissimilarity coefficients between all pairs of individuals in the data. The algorithms are the processes by which a hierarchical clustering technique technically maps a set of observed similarity/dissimilarity coefficients to a new set of similarity/dissimilarity coefficients.

The purpose of this paper is to use average linkage method to cluster the state capitals and the Federal Capital of Nigeria, and to identify the rate of misclassification and the apparent error rate, if any, in carrying out the grouping.

The Thematic Committee in 2001officially grouped the Nigeria states into six geo-political zones based on linguistic affinity, contiguity and cultural affiliation. The zones are; the North-East North-West, North-central, South-East, South-West and South- South. The chairman constitution review committee, Ekweremadu (2010), said that one of the reasons why Nigerians wanted the 1999 constitution amended was because of the structural imbalances in the composition of the states among the six geo-political zones of Nigeria. For instance, while the South-East zone has only five states, North-Central and North-West have seven states each, while

the other three zones have six states each. However, it is not clear how strong these zonings are supported by the traditional way by which Nigerian states have been grouped.

Essentially, cluster analysis exhibits the properties of external isolation and internal cohesion (Cormack, 1971). External isolation requires that objects in one cluster should be separated from objects in another cluster by fairly empty areas of space. Internal cohesion requires that objects within the same cluster should be similar to each other, at least within the local metric. Umeh *et al* (2003) clustered eighteen languages in the old Cross-River State of Nigeria using single linkage, complete linkage and average linkage methods. Their purpose of clustering was to establish similarities among the languages. Comparing the dendrograms derived from these methods, it was found that the average linkage method had a configuration very much similar to complete linkage method. Average linkage method showed the highest number of clusters as most of the languages formed single clusters. Although, the methods merged all the languages into one big cluster, they concluded that single linkage method performed best in terms of clustering faster than other methods.

The rest of the paper is arranged as follows: Section two gives a brief on the hierarchical clustering technique. Section three dwelt on *Pvclust* which is an add-on statistic in the R software and the methods of the paper. Section four is on the results and discussion, while section five is on conclusion.

2.1 Distance Measures

The measure of distance between any n objects, for hierarchical algorithms is given by the Euclidean distance function d_{ij} ;(Bryan, 2005) and is defined by

$$d_{ij} = \{\sum_{k=1}^{p}(y_{ik} - y_{jk})^2 .$$

Where y_{ik} is the value of the variable $y_k$for object i and $y_{jk}$is the same variable for object j.

2.2 Clustering Method

In this paper, the average linkage method was used for the clustering. This method uses the average distance between pairs of objects, say, in cluster A and cluster B. Two groups merge if the average distance between them is small enough. Breiman *et al* (1984), instead of using a minimum and maximum measure, used the average linkage method to calculate the distance between two clusters using the average of the dissimilarities in each cluster as below:

$$d(A,B) = \frac{1}{n_A n_B}\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} dist(y_{Ai}, y_{Bj})$$

At each step, we join the two clusters with the smallest distance. Each individual city is assigned to the group with the nearest average.

2.3 Correlation Method

For data expressed as (nxp) matrix or data frame, we assume that the data is n observations of p objects, which are to be clustered. The i^{th} row vector corresponds to the i^{th} observation of these objects and the j^{th} column vector corresponds to a sample of j^{th} object with size n. There are several methods to measure the dissimilarities between objects. For data matrix $X = \{x_{ij}\}$, the default is "correlation" method:

$$d_2 = 1 - \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x})(x_{ik}-\overline{x_k})}{\{(\sum_{i=1}^{n}(x_{ij}-\bar{x})^2(\sum_{i=1}^{n}(x_{ik}-\overline{x_k})^2)\}^{1/2}}$$

2.4 Dendrogram

Dendrogram is a figure that illustrates how agglomeration takes place (Bryan, 2005). The results of hierarchical clustering techniques are presented in the form of dendrogram (Sneath and Sokal, 1973) or a tree diagram (Hartigan, 1975). Dendrograms are constructed from (n x n) distance matrix. The first step in the construction of a dendrogram, in general, is to arrange the individuals into a hierarchical order such that individuals with the highest mutual similarity are placed together. Then groups or clusters of objects are associated with other groups, which they most closely resemble, and so on until all of the individuals have been placed into a complete classification scheme. However, cluster arising from different distance measures and clustering techniques applied to the same data will differ according to algorithms adopted for distance measures.

3.0 Materials and Methods

The *Pvclust* which is an add-on package for a statistical software R was used to assess the uncertainty in hierarchical cluster analysis in data on distance between Abuja and other states capital in Nigeria. It was also used to calculate probability values (*p-values*) for each cluster using bootstrap resampling techniques. The two types of *p-values* that were obtained are the: approximately unbiased (AU) *p-value* and bootstrap probability (BP) value. Multiscale bootstrap resampling technique was used for the calculation of AU *p-value*, which has superiority in bias over BP-value calculated by the ordinary bootstrap resampling, (Suzuki and Shimodaira, 2006). The algorithms of multiscale bootstrap are:

(i) Generate bootstrap samples for each sample size.
(ii) Apply hierarchical clustering to each bootstrap sample to obtain the sets of bootstrap replications of dendrograms.
(iii) Compute bootstrap probability for each sample size.
(iv) Using values of bootstrap probabilities, estimate the *p*-value by fitting a theoretical equation to them. The estimated *p-value* is called AU (approximately unbiased) value.

3.1 Basic Assumptions

(i) The distance is assumed to be zero between a capital city and itself
(ii) The distance from city i, to city j is the same as the distance from city j to city i

4.0 Results and Discussion

We conducted hierarchical cluster analysis with multiscale bootstrap with number of bootstrap 1000, using average linkage method and correlation based dissimilarity matrix for

testing at first, and then used number of bootstrap 10000 for smaller error as in table 3.3 and 3.4 respectively.

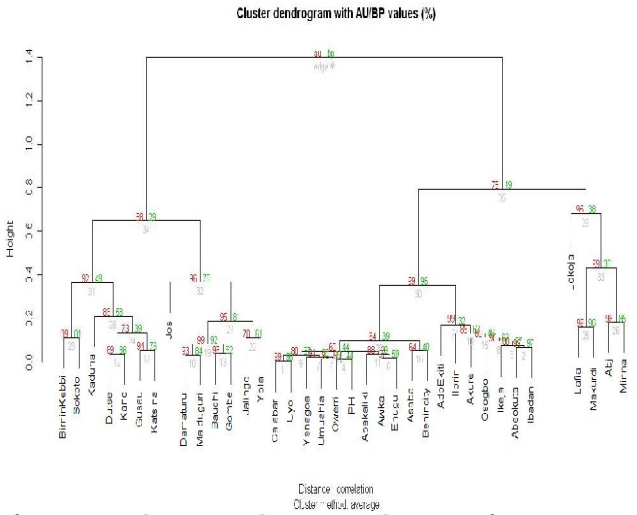

fig. 3.1: Dendrogram with AU/BP values (%) of 37 state capitals in Nigeria

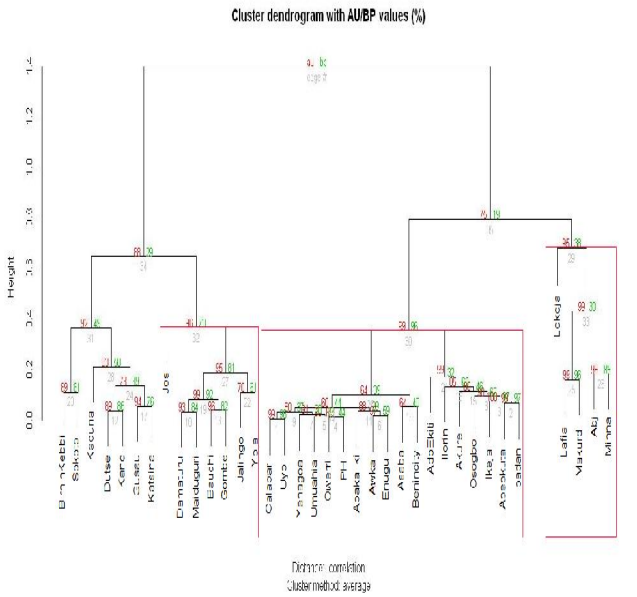

Fig. 3.2: Highlighted dendrogram with AU/BP values (%) of 37 state capitals in Nigeria

Figures (3.1) and (3.2) are the dendrograms showing the average linkage method of hierarchical clustering with *p-values* of 37 state capitals of Nigeria. Red values at left branch are AU (Approximated Unbiased) *p-values*; green values at right branch are BP (Bootstrap Probability) values in percentage, and cluster labels (bottom). Labels at leaves are classification by state capitals. Cluster with AU $\geq$ 0.95 is indicated by the rectangle and the hypothesis that "the cluster does not exist" is rejected with significance level 0.05. The two figures show how the thirty seven state capitals are grouped into six main clusters as can be seen in table 3.1 below. The AU *p-values* are shown in parentheses.

Table 3.1: Six cluster formation of 37 state capitals of Nigeria

Cluster I	Cluster II	Cluster III	Cluster IV	Cluster V	Cluster VI
Birnin Kebbi (0.89)	Bauchi (0.95)	Calabar (0.99)	Abakaliki (0.88)	Abeokuta (0.99)	Abuja (0.96)
Dutse (0.89)	Damaturu (0.98)	Owerri (0.69)	Asaba (0.64)	Ado Ekiti (0.99)	Lafia (0.99)
Gusau (0.94)	Gombe (0.94)	PH (0.60)	Awka (0.97)	Akure (0.99)	Lokoja (0.95)
Kaduna (0.83)	Jalingo (0.99)	Umuahia (0.90)	Benin City (0.99)	Ibadan (0.99)	Makurdi (0.99)
Kano (0.73)	Jos (0.95)	Uyo (0.99)	Enugu (0.64)	Ikeja (1.00)	Minna (0.96)
Kastina (0.83)	Maiduguri (0.99)	Yenago (0.91)		Ilorin (0.85)	
Sokoto (0.92)	Yola (0.99)			Oshogbo (0.99)	

However, the AU *p-values* themselves include sampling error, since they are also computed by a limited number of bootstrap samples. The standard errors of AU *p-values* for all the state capitals can graphically be checked as in fig3.3 below.

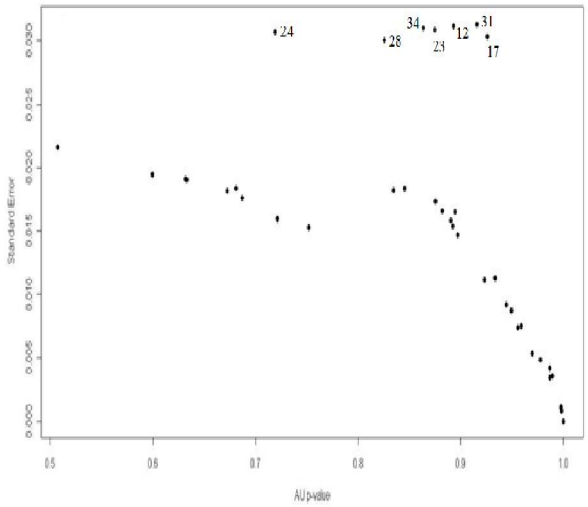

fig. 3.3: *P-values* against standard errors plot of the nboot = 1000

Figure (3.3) show some state capitals whose standard errors are extremely high (larger than 0.030 with AU *p-values* around 0.894, 0.944, 0.886, 0.730, 0.826, 0.830 and 0.920 respectively). The estimated values of these state capitals are given in table 3.3 below:

Table 3.3: *P-values* and Standard Error on Edges of the nboot = 1000

Cluster Label	**au**	**Bp**	**se.au**	**se.bp**	**v**	**c**	**pchi**	**CI for AU *P- value***
12(Dutse)	0.894	0.861	0.611	0.004	-1.166	0.083	0.034	-0.342 and 2.130
17(Gusau)	0.944	0.760	0.273	0.005	-1.149	0.443	0.452	0.390 and 1.498
23(B/Kebbi)	0.886	0.804	0.430	0.005	-0.673	0.374	0.247	0.016 and 1.756
24(Kano)	0.730	0.389	0.391	0.005	-0.159	0.440	0.509	-0.052 and 1.512
28(Kaduna)	0.826	0.576	0.314	0.005	-0.505	0.372	0.098	0.198 and 1.454
31(Katsina)	0.830	0.488	0.251	0.005	-0.698	0.729	0.257	0.328 and 1.332
34(Sokoto)	0.920	0.291	0.196	0.005	-0.301	0.851	0.375	0.528 and 1.312

Table 3.3 show estimates on edges of the state capitals with higher standard errors. For example, Dutse with cluster label 12 has AU *p-value* estimated at 0.894, with SE (standard error) 0.618. By an analogue of standard normal theory, the true AU *p-value* is estimated to exist in between (AU - 2 * SE) and (AU + 2 * SE). From this inference, the true AU *p-value* is between -0.342 and 2.130. However, AU is defined to be between 0 and 1, so this inference fails for such large SE values.

To decrease the standard errors of AU *p-values* of these state capitals, the number of bootstrap (nboot) is be increased to 1000*2. Similar analysis was conducted with nboot = 10000 and the following results were obtained:

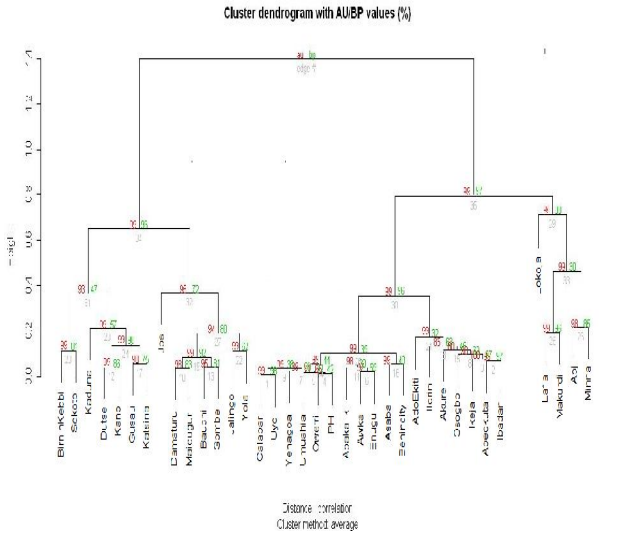

Figure 3.3: Dendrogram with AU/BP values (%) of 37 state capitals in Nigeria

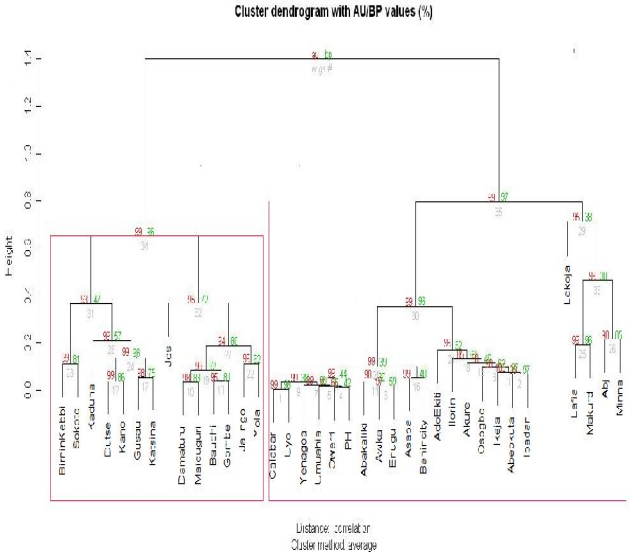

Figure 3.4: Highlighted dendrogram with AU/BP values (%) of 37 state capitals in Nigeria

Table 3.3: Six cluster formation of 37 state capitals in Nigeria of the nboot = 10000

Cluster I	Cluster II	Cluster III	Cluster IV	Cluster V	Cluster VI
Birnin-Kebbi	Bauchi (0.95)	Calabar (0.99)	Abakaliki (0.99)	Abeokuta (0.99)	Abuja (0.96)
Dutse (0.99)	Damaturu (0.98)	Owerri (0.69)	Asaba (0.99)	AdoEkiti (0.99)	Lafia (0.99)
Gusau (0.96)	Gombe (0.94)	PH (0.69)	Awka (0.97)	Akure (0.99)	Lokoja (0.95)
Kaduna (0.99)	Jalingo (0.99)	Umuahia (0.91)	BeninCity (0.99)	Ibadan (0.99)	Makurdi (0.99)
Kano (0.99)	Jos (0.95)	Uyo (0.99)	Enugu (0.97)	Ikeja (1.00)	Minna (0.96)
Kastina (0.93)	Maiduguri (0.99)	Yenagoa (0.99)		Ilorin (0.85)	
Sokoto (0.99)	Yola (0.99)			Oshogbo (0.99)	

Table 3.4: *P-values* and Standard Error on Edges of the nboot = 10000

Cluster label	au	bp	se.au	se.bp	V	c	pchi	CI for AU *P-value*
12(Dutse)	0.986	0.861	0.004	0.001	-1.252	0.262	0.181	0.978 and 0.994
17(Gussau)	0.975	0.744	0.007	0.002	-1.285	0.603	0.167	0.961 and 0.989
23B/Kebbi)	0.993	0.814	0.006	0.001	-1.144	0.051	0.008	0.981 and 1.005
24(Kano)	0.989	0.964	0.009	0.002	-0.046	0.205	0.000	0.971 and 1.007
28(Kaduna)	0.991	0.576	0.009	0.002	-0.027	0.297	0.000	0.973 and 1.009
31(Katsina)	0.932	0.484	0.013	0.002	-0.724	0.764	0.124	0.906 and 0.958
34(Sokoto)	0.984	0.958	0.005	0.002	-0.381	0.913	0.167	0.974 and 0.994

In studying the city capital, Dutse, with cluster label 12. By applying 2 * SE rule again, the "true" AU p-value exists between 0.978 and 0.994. Most commonly we adopt 95% significance level, which seems to be stable enough to conclude the existence of these clusters.

The rate of misclassification can be obtained by tabulating the real six geo-political zones against the clusters determined by kmeans:

Table 3.5: Classification Table for the six cluster formations of 37 state capitals in Nigeria

Six Geo-political Zones of Nigeria	Number of State Capitals	Predicted Group					
		I	II	III	IV	V	VI
North West	**7**	**7**	**0**	**0**	**0**	**0**	**0**
North East	**7**	**0**	**7**	**0**	**0**	**0**	**0**
South South	**6**	**0**	**0**	**6**	**0**	**0**	**0**
South East	**5**	**0**	**0**	**0**	**5**	**0**	**0**
South West	**7**	**0**	**0**	**0**	**0**	**6**	**1**
North Central	**5**	**0**	**0**	**0**	**0**	**0**	**5**

$$Apparent\ correct\ classification\ rate = \frac{7+7+6+5+6+5}{37} = \frac{36}{37} = 0.97$$

$$Apparent\ error\ rate = 1 - 0.973 = 0.027$$

The result show that Clusters I, II, III, and IV, associated perfectly, that is, no any state capital city is misclassified in the four clusters. Cluster V has one misclassification (1 out of 7 cities is allocated to cluster VI), while cluster VI matched perfectly but one city of cluster V is included. This is impressive, given that there are little obvious overlaps in the bootstrap samples for each sample size data.

The recommended six cluster formation of 37 state capitals of Nigeria is given in the table 3.6 below. *P-value* of each cluster is a value between 0 and 1, which indicates how accurate the cluster is.

Table 3.6: Cluster formation of 37 state capitals of Nigeria with AU *p*-value 0.99 for all the Clusters

Cluster I	Cluster II	Cluster III	Cluster IV	Cluster V	Cluster VI
BirninKebbi	Bauchi	Calabar	Abakaliki	Abeokuta	Abuja
Dutse	Damaturu	Owerri	Asaba	AdoEkiti	Lafia
Gusau	Gombe	PH	Awka	Akure	Lokoja
Kaduna	Jalingo	Umuahia	Benin City	Ibadan	Makurdi
Kano	Jos	Uyo	Enugu	Ikeja	Minna
Kastina	Maiduguri	Yenagoa		Ilorin	
Sokoto	Yola			Oshogbo	

Conclusion

This paper has assessed the uncertainty in results of average linkage method and correlation based dissimilarity matrix of hierarchical cluster analysis of state capitals in Nigeria. We conducted hierarchical cluster analysis with multiscale bootstrap with number of bootstrap 1000, for testing at first, and then used number of bootstrap 10000 for smaller error. We are more confident of the estimation accuracy of AU *p-values* as well as the existence of highlighted clusters. The study show six cluster formation for Nigeria with seven state capitals in clusters I, II, and V; with six state capitals in cluster III, and five state capitals in clusters IV and VI. The states in these classifications are different from the states in the geo-political zones of Nigeria.

REFERENCES

1. Breiman *et al.* (1984), Cluster Analysis and Multidimensional Scaling. The Classification and Regression Tree (CART) Methodology (pp.516-528).
2. Bryan. F.J.M. (2005). Multivariate Statistical Methods. A Primer; 3rd edtn.,Chapman and Hall/CRC, USA.
3. Cormack, R. M. (1971), A review of classification, Journal of Royal Statistical Society(Series A), 134, 321-367.
4. Ekweremadu, I. (2010), Nigeria: Amended Constitution-the Gains, the Knocks, theExpectations, http://allafrica.com/stories/201007200264 .html.,
5. http://en.wikipedia.org/wiki/cluster_analysis
6. Hartigan, J. A. (1975). Clustering Algorithms. New York: Wiley.
7. Sneath, P. H. A. and R. R. Sokal (1973). Numerical Taxonomy. San Francisco: Freeman.
8. Suzuki, R. and Shimodaira, H. (2006) Pvclust: an R package for hierarchical clustering with p- values.
9. Suzuki, R. and Shimodaira, H. (2006), "An application of multiscale bootstrap resampling to hierarchical clustering of microarray: How accurate are these clusters?" In proceedings by the fifteenth International Conference on Genome Informatics (GIW 2004), p. po34.
10. Thematic Committee (2001). Sustainable urban development and good governance in Nigeria.
11. Umeh, E. U. *et al* (2003), A paper presented at the 27th Annual Conference of Nigeria Statistical Association (NSA).