Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 3159-0040
Vol. 2 Issue 10, October - 2015

Clustering Using Principal Component Analysis As An Input Tool

Okeke, Evelyn Nkiruka and Okeke, Joseph Uchenna
Department of Mathematics and Statistics, Federal University Wukari, Nigeria.
Corresponding author: evelyn70ng@yahoo.com

***Abstract*—This study is on the application of principal component analysis to clustering as an approach to reducing the space of the data. In this study two real data sets (on CO_2 emission and 2015 Nigeria's presidential election) were used for comparison. The result of the study revealed that clusters from transformed data (PC data) are more visible and follows previously known grouping than clusters from the original data.**

Keywords—Cluster analysis, Relevant variables, Hierarchical clustering, Single linkage clustering, Principal component analysis, Principal components, Dendrogram, Cluster.

1.0 Introduction

The term cluster analysis is a statistical method for grouping objects of similar kind into respective categories. Indentifying groups of individuals or objects that are similar to each other but different from individual in other groups can be intellectually satisfying, profitable, or sometimes both. A general question facing researchers in many areas of inquiry is how to organize observed data into meaningful pattern. The essential concern of cluster analysis is to find groupings within group of objects, experimental units, or variables (or factor) etc in a way that the degree of association between objects is maximal if they belong to the same group and minimal otherwise. Given the above description, cluster analysis can be used to discover structure in the data without explaining why they exist.

There is countless number of examples in which clustering plays an important role. Using type of fruits as an example by studying the mineral content of different types of fruits you can be able to form group of fruit thereby providing a substitute to each type when it is out of season. By studying different dialects of different people, one can be able to trace their ancestral homes. Based on scores on psychological inventories, you can cluster patients into subgroups that have similar response patterns. This may help you in targeting appropriate treatment and study typologies of disease.

Although both cluster analysis and discriminant analysis classify objects (or cases) into categories, discriminant analysis requires you to know group membership for the cases used to derive the classification rule. The goal of cluster analysis is to identify the actual group in which an object belongs. One of the problems of cluster analysis is on how to distinguish between the relevant and irrelevant variable to be included in the study.

In this paper principal component analysis was used as an input tool to cluster analysis in reducing the dimension of the data thus making the cluster result easy to explore and visualize. This work is divided into sections; section one is on the introduction and description of cluster analysis, section two discussed the principal component analysis, section three is on data and its description, section four is about data analyses and the results, while section five is on the conclusion of the work.

1.1 Relevant Variables in Cluster Analysis

Issues that need consideration prior to caring out any cluster analysis include the following: appropriate scaling or weighting of the variables or transformation of them; measures of proximity, or metrics to use as indicator of closeness among the items to be clustered. Choice made at this stage can have a determining influence on the outputs of the subsequent analysis. Cluster sought should be scale invariant. The nature of the data, as well as the type of clustering one wishes to use in the specific situation will influence the choice of the input. Choice made at input stage can have determining influence on the result of the analysis.

Selecting the variables to be included in the cluster analysis must be done with regard to both theoretical concept and practical considerations. Variable to be included will be only those variable that

- Characterized the object being clustered
- Relate specifically to the objective of the cluster analysis

Cluster analysis has no mechanism for differentiating between relevant and irrelevant variable. Therefore the choice of variable included in a cluster analysis must be underpinned by conceptual considerations. This is very important because the cluster formed can be very dependent on the variable included (Everitt, Londau, and Leese 2001). The inclusion of an irrelevant variable increases the chance that outlier will be created on these variables which will have a substantive effect on the results. Thus one should not include variable indiscriminately, but instead choose the variable with the research objective as the criterion for selection. In practical vein, cluster analysis can be drastically affected by the inclusion of one or two inappropriate or undifferentiated variables. The researcher is always encourage to examine the results and eliminate variables that are not distinctive (i.e. that do not differ significantly) across the derived clusters. This procedure allows for the cluster

techniques to maximally define clusters based only on those variables exhibiting difference across the object.

1.2 Clustering Methods

In clustering there are numerous ways you can sort cases into groups. The choice of a method depends on, among other things, the size of the data file.

If you have a large data file (even, 1000 cases is large for clustering) or a mixture of continuous and categorical variable, you should use the SPSS two-step procedure. In two-step cluster, to make a large problem tractable, in the first step, cases are assigned to "preclusters". In the second step, the preclusters are clustered using the hierarchical cluster algorithm. You can specify the number of clusters you want or let the algorithm decide based on preselected criteria.

If you have a small data set and wants to easily examine solution with increasing number of clusters, you may use hierarchical clustering. For hierarchical clustering, you choose a statistic that quantifies how far apart (or similar) two cases are. Then you select a method for forming the groups. Because you can have as many clusters as you do cases (not a useful solution), your last step is to determine how many clusters you need to represent your data. You do this by looking at how similar clusters are when you create additional cluster or collapse existing ones.

Hierarchical methods includes

Single linkage
Complete linkage
Average linkage
Centroid linkage
Others

In single linkage clustering, the dissimilarity between x and y is the smallest dissimilarity between two points in opposite group, that is,

$\min\{d(x,y): x \in G_1, y \in G_2\}$

In complete linkage clustering dissimilarity between x and y is the largest dissimilarity between two points in opposite group, that is,

$\max\{d(x,y): x \in G_1, y \in G_2\}$

In average linkage clustering dissimilarity between x and y is average dissimilarity over all points in opposite group, that is,

$\frac{1}{|G_1|.|G_2|}\sum_{x\in G_1}\sum_{y\in G_2} d(x,y)$

(WIKIPEDIA 2015)

In centroid linkage the distance between two groups is the Euclidean distance between their centroids, that is,

$d(x,y) = \|\bar{x} - \bar{y}\|^2$

If you know the number of clusters you want to form and you have a moderately sized data set, you use k-mean clustering, In k-mean clustering, you select the number of clusters you want. The algorithm iteratively estimates the cluster means and assigns each case to the cluster for which its distance to the cluster mean is the smallest.

1.3 Partitioning

As an alternative both to hierarchical and to overlapping clustering, partitioning approaches assign each object to exactly one cluster. A generic description of the objective is to maximize similarity/cohesiveness/homogeneity within each cluster while maximizing heterogeneity among clusters.

1.4 Measures of Similarity /Dissimilarity

Distance measures are the most commonly use measures of similarity between objects. Most of the analytical techniques for assessing distance are particularly sensitive to outliers. Screening for outliers is advisable.

Different distance measures leads to different cluster solution. Thus it is advisable to use several measures and compare the results to the theoretical or known patterns.

When the variables have different units, one should standardize the data before running the cluster. Standardization is particularly advisable when the range of one variable is much larger than that of others.

When the variables are intercorrelated (either positively or negatively), the Mahalanobis distance measure is likely to be the most appropriate because it adjusts for intercorrelation and weights all variables equally.

1.41 Euclidean Distance

Euclidean distance is the most common use of distance. In most cases when people talk about distance, they will refer to Euclidean distance. The Euclidean distance between points x and y is the line segment connecting them $\overline{xy}$. In Cartesian coordinates, if $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ are two points in Euclidean n-space, then the distance (d) from x to y or from y to x is given by the Pythagorean formula

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - y_{jk})^2}$$
$$= \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2} \quad (1)$$

1.42 Mahalanobis Distance

To obtain a useful distance measure in a multivariate setting, we must consider not only the variances of the variables but also their covariance or correlations. The simple Euclidean distance between y and $\bar{y}$, $(y_i - \bar{y})'(y_i - \bar{y})$ is not useful because there is no adjustment for the variance or the covariance. For a statistical distance, we standardize by inserting inverse of the covariance matrix:

$$d^2 = \big((y_i - \bar{y})'\big)S^{-1}\big((y_i - \bar{y})\big) \quad (2)$$

The (squared) distance d^2 between two vectors were first proposed by Mahalanobis (1936) and are often referred to as Mahalanobis distances. If a random variable has a larger variance than another, it receives relatively less weight in a Mahalanobis distance. Similarly, two highly correlated variables do not contribute as much as two variables that are less correlated. In essence, then, the use of the inverse of the covariance matrix in a Mahalanobis distance (involving random variables) has the effect of (1)

standardizing all variables to the same variance and (2) eliminating correlations.

1.43 Measures of Association or Dissimilarity coefficients

If we consider two binary-valued vectors x and y then the element –by-element matches are of four types labeled a, b, c, and d as in fig. 1

x \ y	1	0
1	a	b
0	c	d

Fig. 1: The four possible pattern resulting from matching elements of two binary-valued vectors x and y.

For example, if $x = (1, 1, 0, 0)$ and $y = (1, 0, 1, 0)$, then the first entries $(1, 1)$ in each vector are an a-type pair, the second $(1, 0)$ are a b-type pair, etc. An endless number of coefficients of agreement can be written as a function of those four types; for example, Pearson product moment correlation is given by

$$r = \frac{(ad - bc)}{[(a+b)(a+c)(b+d)(c+d)]^{0.5}} \qquad (3)$$

.

Cheetham and Hazel (1969) were among the first to catalogue the various coefficients published and based on the format of fig. 1.1, and their list had less than 24 entries.

2.0 Principal Component Analysis

Principal component analysis is a statistical tool that finds the underlying structure in the data. It finds the direction of most variance in a set of data, that is, the direction where the data is most spread out. This tool brings out strong patterns in a data set and makes data easy to explore and visualize.

Principal component analysis is a variable reduction procedure. This is useful when you have obtained data on a large number of variables or factors and believe that there is some redundancy in those variables. In this case, redundancy means that some of the variables are correlated with one another, possibly because they are measuring the same construct. Because of this redundancy, you believe that it should be possible to reduce the observed variables into a smaller number of principal components (artificial variables) that will account for most of the variance in the observed variables.

It is true that classical PCA is a one sample techniques applied to data with no groupings among observations and no partitioning of variable into subset Y and X for instance (Rencher 1995) but it has been found useful in multiple regression when computational or statistical problems arise in the presence of severe multicollinearity in the set of independent variables $X = X_1, \dots, X_p$ (Hadu and Linq 1989) and where there are too many independent variables relative to the number of observations (Rencher 1995); discriminant analysis with small sample situation, perhaps with fewer observations than variables (Kahirsaga, Kocherkots, and Kochecloakata 1990); cluster analysis when Euclidean distance between observation in p-dimensional space are closely approximated by Euclidean distance in space spanned by the first q principal components, provided the first q eigenvalues account for a high percentage of the trace of the covariance matrix; canonical regression when you have so many dependent variables correlating with a set of p independent variables.

2.1 Mathematical Details of Principal Component Analysis

Consider forming a scalar variate Y as a linear combination of the original variate X_i

$$Y = a_1 X_1 + \cdots + a_p X_1 = a_i{}' X_i \qquad (4)$$

where $\boldsymbol{a_i} = \boldsymbol{a_1} + \cdots + \boldsymbol{a_p}$ is a p-dimensional vector.

The mean of Y is

$$\bar{\mathrm{Y}} = \mathrm{a}_1 \bar{\mathrm{X}}_1 + \cdots + \mathrm{a}_2 \bar{\mathrm{X}}_2 = \mathrm{a}_\mathrm{i} \bar{\mathrm{X}}_\mathrm{i} \qquad (5)$$

Consider the variance of Y

$$S^2 = \frac{1}{\mathrm{n}-1} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y}_\mathrm{i} - \bar{\mathrm{Y}})^2 \qquad (6)$$

since

$$y_i - \bar{Y} = a_i{}' x_i - a_i{}' \bar{X} = a_i{}' (x_i - \bar{X}) \qquad (7)$$

and the transpose of a scalar is equal to that scalar

$$\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y}_\mathrm{i} - \bar{\mathrm{Y}})^2 = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{a}'(x_i - \bar{X})(x_i - \bar{X})a \qquad (8)$$

Since **a** is a constant over i, hence

$$S^2 = \frac{1}{\mathrm{n}-1} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y}_\mathrm{i} - \bar{\mathrm{Y}})^2 = \mathrm{a}' \left[\frac{1}{\mathrm{n}-1} \sum_{\mathrm{i}=1}^{\mathrm{n}} (x_i - \bar{X})(x_i - \bar{X}) \right] \mathrm{a} \qquad (9)$$

$$= a' S a \qquad (10)$$

By multiplying out the vectors, the quadratic form $a'Sa$ can be expressed in terms of the elements of **a** and **S** as

$$a_i{}' a_j S_{ij}$$

Algebraically, we can define the first principal component to be the linear combination $Y_1 = a_1{}' X$ of the original variables X subject to the constraint that $a_1{}' a_1 = 1$, where a_1 is the corresponding eigenvector of the first eigenvalue of S_{ij}. We therefore require finding the vector a_1 that satisfies this condition. The conditions for vector a_1 to maximize $a_1{}' S a_1$ subject to the constraint $a_1{}' a_1 = 1$ are precisely the same as those for the vector a_1 to maximize

$$\mathrm{a}_1' \mathrm{S} \mathrm{a}_1 - \lambda_1 (\mathrm{a}_1{}' \mathrm{a}_1 - \mathrm{k}) \qquad (11)$$

The standard procedure for maximizing a function of several variables subject to one or more constraints is by the method of Lagrange multiplier. With just one constraint, the method uses the fact that the stationary point of differentiable function of p variables, say $f(X_1, \dots, X_p)$, subject to constraint $g(X_1, \dots, X_p) = C$ are such that there exist a number Λ, called *Lagrange multiplier*, such that

$$\frac{df}{dx_i} = \frac{\lambda df}{dx_i} = 0 \qquad for\ i = 1, \dots, p \qquad (12)$$

at stationary points.
These p equations, together with the constraints, are sufficient to determine the coordinate of the stationary points. Further investigation to see if a stationary point is maximum, minimum or saddle points is made using a new function, $L(X)$ such that

$$L(X) = f(X) - [g(X) - C] \qquad (13)$$

where the term in the square bracket is of course zero. The derivative of $L(X)$ is

$$\frac{\boldsymbol{d}\, L(X)}{\boldsymbol{dx}} = \frac{df(x)}{dx_i} \qquad (14)$$

Using the form of $L(X)$, and letting

$$L(a_1) = a_1' S a_1 - \lambda_1 (a_1{}' a_1 - 1) \qquad (15)$$
$$= \sum_{i=1}^{p}\sum_{i=1}^{p} a_{1i}{}' a_{1j} S_{ij} - \lambda_1 \left(\sum_{i=1}^{p} a_{1i}^2 - 1 \right)$$

where $a_{1i} = (a_{11}, \dots, a_{1p})$. Then

$$\frac{dL(a)}{da_{1k}} = 2\sum_{j=1}^{p} S_{kj} a_{1j} - 2\lambda_1 a_{1k} \qquad where\ k = 1, \dots, p \qquad (16)$$

To find the vector a_{1i}' that maximizes $L(a)$
We set $\frac{dL(a)}{da_{1k}} = 0$ for all k and solve the resulting set of simultaneous equations.
Now

$$\sum_{j=1}^{p} S_{kj} a_{1j} = \lambda_1 a_{1k} \qquad (17)$$

The left hand side is the *kth* element of Sa_1, while the right hand side is the kth element of $\lambda_1 a_{1k}$.
Thus when all k equations are treated simultaneously, it follows that the maximizing value of a_1 must satisfy

$$Sa_1 = \lambda_1 a_1 \qquad (18)$$

That is,

$$(S - \lambda_1 \mathrm{I}) a_1 = 0$$

This is a homogeneous set of $\boldsymbol{p}$ equations in $\boldsymbol{p}$ unknowns. For non-trivial solution $a \neq 0$

$$|S - \lambda_1 \mathrm{I}| = 0 \qquad (19)$$

λ_1 is the first eigenvalue of S and the solution a_1 is its corresponding eigenvector normalized so that $a_1{}' a_1 = 1$.
Now we consider the second principal component. We look for a second linear combination

$$Y_2 = a_i{}' X_i \qquad (20)$$

of the original variable X subject to constraint that and $a_2' a_2 = \sum_{i=1}^{p} a_{2i}^2 = 1$, where $a_2 = a_{21}, \dots, a_{2p}$. In addition, this line (Y_2) must be orthogonal to the one defining the first principal component, the condition for which is that $a_2' a_1 = a_1' a_2 = \sum_{i=1}^{p} a_{1i}' a_{2i} = 0$. The variance of Y_2

$$var(Y_2) = a_2' S a_2 \qquad (21)$$

Maximizing this variance subject to the constraints above will involve two Lagrange multipliers (λ_2 and m) and thus we must require to maximize

$$L(a_2) = a_2' S a_2 - \lambda_2 (a_2' a_2 - 1) - m(a_2' a_1) \qquad (22)$$
$$= \sum_{i=1}^{p}\sum_{j=1}^{p} a_{2i} a_{2j} S_{ij} - \lambda_2 \left(\sum_{i=1}^{p} a_{2i}^2 - 1 \right) - m \left(\sum_{i=1}^{p} a_{1i} a_{2i} \right)$$

Thus,

$$\frac{dL(a_2)}{da_{2k}} = 2\sum_{i=1}^{p} S_{kj} a_{2j} - 2\lambda_2 a_{2k} - m a_{1k} \qquad where\ k = 1, \dots, p \qquad (23)$$

Setting $\frac{dL(a_2)}{da_{2k}} = 0$ for all k leads to the equation

$$(S - \lambda_2 \mathrm{I}) a_2 = \frac{1}{2} m a_1 \qquad (24)$$

By multiplying (24) by a_1', and recalling that $a_1{}' a_1 = 1$, while $a_1{}' a_2 = 0$ we see that

$$a_1' S a_2 = \frac{1}{2} m \qquad (25)$$

By also multiplying (24) by $a_2{}'$ and remembering that $a_2{}' a_1 = 0$ however yield $a_1' S a_2 = 0$. Since $a_1' S a_2$ is a scalar quantity and S is a symmetric matrix then $a_1' S a_2 = a_2' S a_1 = 0$. Substituting in (25) thus yields $m = 0$, and from (24) we see that the coefficient of the second principal component also satisfy $(S - \lambda_2 \mathrm{I}) a_2 = 0$. The fact that the variance of the second principal component must be maximum after the first component has been accounted for shows that the coefficient of the second principal component (Y_2) are giving by the elements of the eigenvector a_2 correspond to the second largest eigenvalue λ_2 of S. Continuing in this manner, the element of a_j turn out to be the eigenvector associated with jth largest eigenvalue λ_j.

3.0 Numerical Application

3.1 Source of Data

The two data sets we used in this study are published data on CO_2 Emission of 2011 and their possible correlates for some countries (United States, China, Russia, India, Japan, Germany, Canada, and United Kingdom) and 2015 Nigeria's presidential election (INEC Nigeria 2015). The possible correlate for CO_2 emission we considered includes GDP, Industrial output, export output, energy consumption and manufacturing output. In the 2015 Nigeria's presidential election the votes of 14 registered political parties for the 36 states of the Federation including the Federal Capital Territory were considered.

4.0 Analyses and Results

Principal component analyses of the original data sets were computed to enable us to reduce the dimension of the data sets. In CO_2 emission data, two principal components (PCs) that account for more than 80 percent of the total variation in the original data were selected. In 2015 Nigeria's presidential election data,

four principal components were selected. The selected principal components from the original data sets were used to transform the original data sets. The cluster analyses of the original data sets and that of the transformed data (PC data) were done using Minitab computer package. The dendrograms of few out of gthe many analyses are shown in the following figures below:

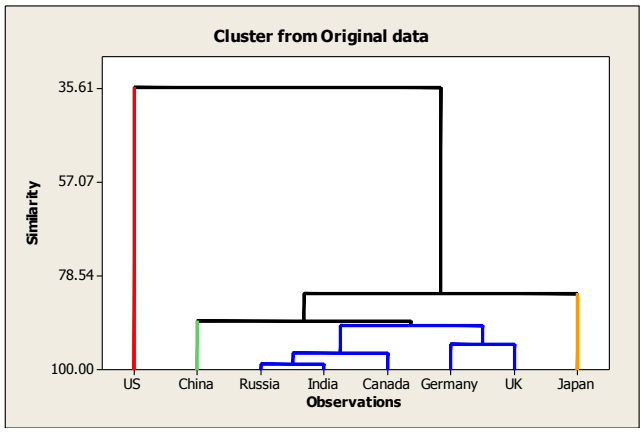

Fig. 2a: Single linkage Cluster of CO_2 Emission (Original data)

In fig. 2a we observed the existence of four clusters thou the clusters seem to overlap. In that dendrogram Russia, India, and Canada fall in one group with similarity rate of 96.29 percent; Germany and UK fall in a group with similarity rate of 94.18 percent; China joins the existing two groups with similarity rate of 89.02 percent while Japan joins them with similarity rate of 82.71 percent. With the rates at which China and Japan joined the clearly observed two groups they are merged to form a group. US is on a separate group and it joins other group at similarity rate of 36.61 percent.

In fig. 2b four non-overlapping clusters were observed. Russia, and India are joined together at similarity rate of 99.03 percent; Canada joins them with similarity rate of 96.58 percent. With this rate at which Canada joins the Russia and India we decided to put Russia, India, and Canada in one group thereby disregarding the group of Russia and India alone. Germany and UK fall in a group with similarity rate of 93.01 percent; China joins the existing two groups with similarity rate of 87.04 percent while Japan joins them with similarity rate of 85 52 percent. These rates of China and Japan made us to group them in one group. US is on a separate group and it joins already existing groups at similarity rate of 35.65 percent.

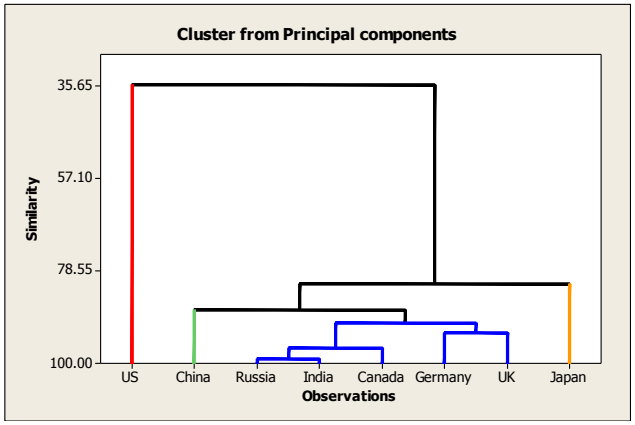

Fig. 2b: Single linkage Cluster of CO_2 Emission (PC data)

Figures 2a and 2b also revealed that among the eight countries we studied that Russia and India are among the countries that produce least CO_2emission. Among those countries US is the highest producers of CO_2 emission.

From fig. 3a the number of groups cannot be clearly stated because the clusters are overlap. But from fig. 3b, five groups (or clusters) were observed. From that dendrogram were observed that Abia, Beyelsa, Ebonyi, Cross River, Anambra, Enugu, and Imo have the same voting pattern and as a result they are merged together to form a group.

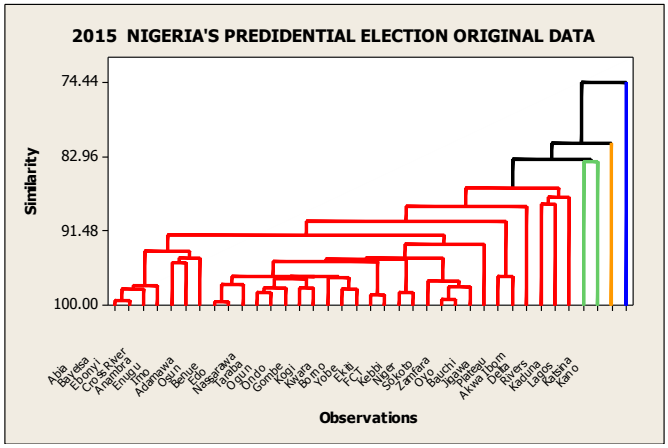

Fig. 3a: Single linkage Cluster of 2015 Nigeria's Presidential Election (Original Data)

Adamawa, Osun, Kogi, Ogun, Benue, Ondo, Lagos, Kwara, Edo, Ekiti, Nasarawa, Taraba, FCT, Oyo, Gombe, and Plateau have similar voting pattern and for this they are grouped together in one cluster. Jigawa and Bauchi have similar voting pattern and they are grouped in one cluster. Kaduna, Sokoto, Niger, Yobe, Zamfara, Kebbi, and Borno have similar voting pattern and for this they formed a group. Kano, Katsina, Rivers, Delta and Akwa Ibom have similar voting pattern and they are grouped together in one cluster. Observe that the states that formed the last cluster are from the geographical regions (South south and North west) from which the two major contenders President Buhari and Ex-president Goodlock came from.

From the five groups (or clusters) we obtained, we discovered that all the South East states are in the same group; all the South West states are in the same group; the North central states apart from Kaduna and Niger fall in the same group.
From figure 3b also we observed that the highest number of votes came from Kano.

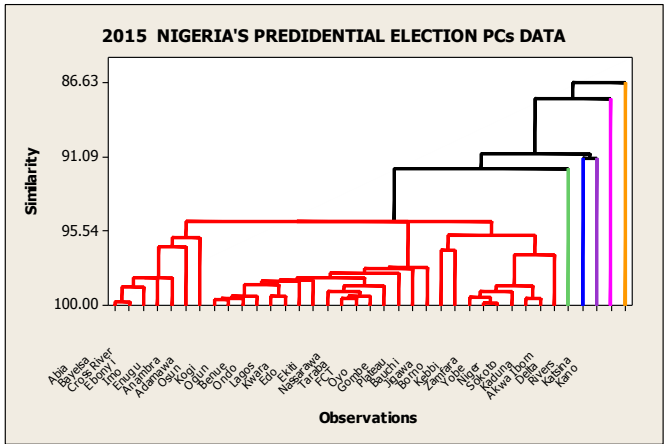

Fig. 3b: Single linkage Cluster of 2015 Nigeria's Presidential Election (PC data)

5.0 Conclusion

From the result of the analyses we conclude that Principal component analysis is very efficient input tool to cluster analysis in grouping objects into cluster(s) because it produces already existing pattern in original data and also gives better display than the original data especially when the number of variables are many. Other hierarchical clustering methods we studied (complete linkage, average linkage, and centroid linkage) show the same result. The Minitab package was chosen for this study because it produces clear dendrogram than other statistical packages available to us.

Reference

[1] Cheetham and Hazel (1969) On the Use of Cluster Analysis in Biogeography Systematic Biology 21(2): 240-242
[2] Everitt, B.S., Londau S., and Leese. M.(2001). Cluster Analysis. Taylor and Francis, United States
[3] Hadi, A. S. and Ling, R. F. (1998). Some cautionary note on the use of principal component regression, Journal of the American Statistical Association, 52:1, 12-19.
[4] Independent National Electoral Commission Nigeria (2015). 2015 Presidential election result
[5] Kshirsaga, A. M., Kocherlokota, S., and Kocherlokota, K. (1990). Classification procedure using principal component analysis and stepwise discriminant function, Communication in Statistics - Theory and Method, 91-109.
[6] Mahalanobis, P. C.(1936). On the generalized distance in statistics, Proceeding of the National Institute of Science of India, pages 49-55.
[7] Rencher,A.C. (1995).Method of multivariate analysis. John Wiley and Sons, Canada 280-281.
[8] WIKIPEDIA (2015). Hierarchical Clustering, 8th July 2015.