Research Journal of Applied Sciences 7 (1): 41-47, 2012
ISSN: 1815-932X

Simultaneous Application of Agglomerative Algorithms on Interval Measures for Better Classification of Crime Data Across the States in Nigeria

[1]Y. Bello, [2]S.U. Gulumbe and [3]S.A. Yelwa
[1]Department of Mathematics and Statistics, Hassan Usman Katsina Polytechnic, Katsina, Nigeria
[2]Department of Mathematics, [3]Department of Geography,
Usmanu Danfodiyo University, Sokoto, Nigeria

Abstract: Agglomerative algorithms have been used in crime analysis to classify criminal activities and to classify areas into higher and lower criminal activities. The researches were mostly applying the algorithms on one particular distance measure. This study identifies the usefulness of applying different algorithms on different interval measures simultaneously for better classifications of crime data across the 36 states in Nigeria using statistical analysis supplemented with Geographical Information Systems (GIS) analysis.

Key words: Agglomerative algorithms, interval measures, crime data, GIS, classification, Sokoto

INTRODUCTION

Understanding the context of crime the where and when of a criminal event is key to understanding how crime can be controlled and prevented (Cahill, 2004). Geographic classifications using Geographic Information System has become increasingly important in law enforcement and crime prevention. Police desire information on crime for the geographic areas in order to precisely target patrols and investigative efforts (Radoff, 1993; Fajemirokun *et al*., 2006). However, this study applies the technique of multivariate cluster analysis for classifying the Nigeria states in to areas of higher and lower crime concentration.

Hierarchical cluster analysis is comprised of agglomerative methods and divisive methods that find clusters of observations within a data set. The divisive methods start with all of the observations in one cluster and then proceeds to split (partition) them into smaller clusters. The agglomerative methods begin with each observation being considered as separate clusters and then proceeds to combine them until all observations belong to one cluster. Four of the better-known algorithms for hierarchical clustering are average linkage, complete linkage, single linkage and Ward's linkage. Agglomerative algorithms are used quite frequently in practice (Lasch *et al*., 2004). Tsiamtsiouri and Panaretos (1999) have applied Complete Linkage, the Centroid and the Ward's Method to analyze Greek crime data. Rencher (2002) has used the US crime data to demonstrate the different agglomerative algorithms. Hardle and Zdenek (2007) have performed Ward algorithm on the Euclidean distances between standardized observations to identify the states of the US with higher and lower crime concentrations. Maydeu-Olivares (1996) has used applied six different clustering algorithms on the Euclidean distances to classify prison inmates and compared the results with the actual qualitative classification. The result of the Ward Method, among others has satisfactorily matched the actual classification. It is in the line of this that this study is written so as to determine whether simultaneous application two of the Agglomerative Algorithms-Complete Linkage and Ward Methods on two interval measures Euclidean and Manhattan distances between standardized observations improves better classification of areas with higher and lower property crime concentrations using crime data in Nigeria. The analysis used the Nigeria crimes against property data set for 2009.

MATERIALS AND METHODS

The crime data for the 36 states and Abuja was collected from the Annual Report of the Nigeria Police Force, 2009 and plotted geographically using geographical coordinates to confirm the proper locations state-wise in a geographical information environment initially. For identification, the 36 state commands and Abuja were identified by their coordinates locations in the six geo-political zones and the zonal numbers of the 12 zonal commands in Nigeria (NPF, 2009). The data were later transferred and subjected to mathematical and statistical analysis. With X denoted as a state, then each state were identified as $_{i,j}X$, i = 1,..., 6 in the order North-West, North-East, North-Central, South-West, South-East and South-South geo-political zones and j = 1, 2,...12 police zonal commands. The property crimes under study are robbery, theft, vehicle theft, burglary, house breakings, store breakings, false pretence and cheating, forgery and arson.

Corresponding Author: Y. Bello, Department of Mathematics and Statistics, Hassan Usman Katsina Polytechnic, Katsina, Nigeria

The value for each crime is converted to crime rate per 100,000 populations of the state and calculated as:

$$\text{Crime rate} = \frac{\text{Number of crime committed}}{\text{Population of the state}} \times 100{,}000 \quad (1)$$

Agglomerative algorithm: Given a data matrix containing multivariate measurements on a large number of individuals (or objects), the objective is to build subgroups for which the observations or objects within each cluster are similar but the cluster are dissimilar to each other according to some appropriate criterion.

The stating point of a cluster analysis is a data matrix χ (n×p) with n measurements (objects) of p variables. The proximity (similarity) among objects is described by a matrix D (n×n):

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix} \quad (2)$$

The matrix D contains measures of similarity among the n objects. If the values d_{ij} are distances, they measure similarity. The greater the distance, the less similar are the objects. The Euclidean distance d_{ij} between two cases, i and j with variable value:

$$x_i = (x_{i1}, x_{i2}, ..., x_{ip})$$

And:

$$x_j = (x_{j1}, x_{j2}, ..., x_{jp})$$

is define by:

$$\begin{aligned} d_{ij} &= \sqrt{(x_i - x_j)'(x_i - x_j)} \\ &= \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ip} - x_{jp})^2} \\ &= \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \end{aligned}$$

For the standardized variables:

$$d_{ij} = \sqrt{\frac{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}{S_{X_k X_k}}} \quad (3)$$

Manhattan distance is the sum of the absolute differences between the values of the item and is define by:

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}| \quad (4)$$

Clusters formed in these methods are compared using the averages of the standardized data set within the clusters obtained to determine the clusters of higher and lower crime concentrations (Hardle and Simar, 2003).

The clustering algorithm starts by calculating the distances between all pairs of observations followed by stepwise agglomeration of close observations into groups. If two objects or clusters say, P and Q are united, one computes the distance between this new group (object) P∪Q and group R using the following updating formula (Lance and Williams, 1967):

$$\begin{aligned} d(R, P \cup Q) = \alpha_1 d(R, P) + \alpha_2 d(R, Q) + \alpha_3 d(P, Q) + \\ \alpha_4 |d(R, P) - d(R, Q)| \end{aligned} \quad (5)$$

where, α_j's are weighting factors that lead to different agglomerative algorithms.

Complete linkage: Clustering uses the farthest (maximum distance between) pair of observations between two groups to determine the similarity of the two clusters. Let d (P, Q) denote the distance between clusters P and Q and n_P and n_Q the number of observations belonging to clusters P and Q, respectively. The complete linkage method of defining the cluster P∪Q is joined and some other cluster, says R is (Hardle and Simar, 2003):

$$d(R, P \cup Q) = \max \{d(R, P) + d(R, Q)\}$$

The Ward clustering algorithm computes the distance between groups according to the Eq. 5. The Ward algorithm does not put together groups with smallest distance. Instead, it joins groups that do not increase a given measure of heterogeneity too much. The aim of the Ward procedure is to unify groups such that the variation inside these groups does not increase too drastically: the resulting groups are as homogenous as possible.

The heterogeneity of group R is measured by the inertia inside the group. This inertia is defined as:

$$I_R = \sum_{i=1}^{n_R} d^2\left(x_i, \overline{x_R}\right) \quad (6)$$

where, $\overline{x_R}$ is the arithmetic average and the n_R the number of observation within group R (Ward, 1963). If the usual Euclidean distance is used then I_R represents the sum of variances of the component of x_i inside group R. When two observations or groups P and Q are joined, the new group P∪Q has a larger inertia $I_{P \cup Q}$ and the corresponding increase of inertia is given:

$$\Delta(P,Q)=\frac{n_P n_Q}{n_P+n_Q}d^2(P,Q) \quad (7)$$

In this case, the Ward algorithm is defined as an algorithm that joins the groups that give the smallest increase in Δ (P, Q). When the new group P∪Q is joined with some other group, says R, the increase of inertia (from Eq. 5) is given by:

$$\Delta(R,P\cup Q)=\frac{n_R+n_P}{n_R+n_P+n_Q}\Delta(R,P)+\frac{n_R+n_Q}{n_R+n_P+n_Q}\Delta(R,Q)-\frac{n_R}{n_R+n_P+n_Q}\Delta(P,Q)$$

where, n_P, n_Q and n_R are the number of observations belonging to clusters P, Q and R, respectively. In Ward Method, $\alpha_4 = 0$.

RESULTS AND DISCUSSION

The aim is to determine a suitable state classification into higher and lower crime clusters using different Hierarchical Methods. The analysis was carried out by performs Wards and Complete Linkage Methods on block and Euclidean distances.

Complete linkage: The resulting dendrogram of the complete linkage performed on the Euclidean distances is shown in Fig. 1 and has suggested splitting the Nigeria crime data set into six clusters. The resulting clusters are shown in Table 1.

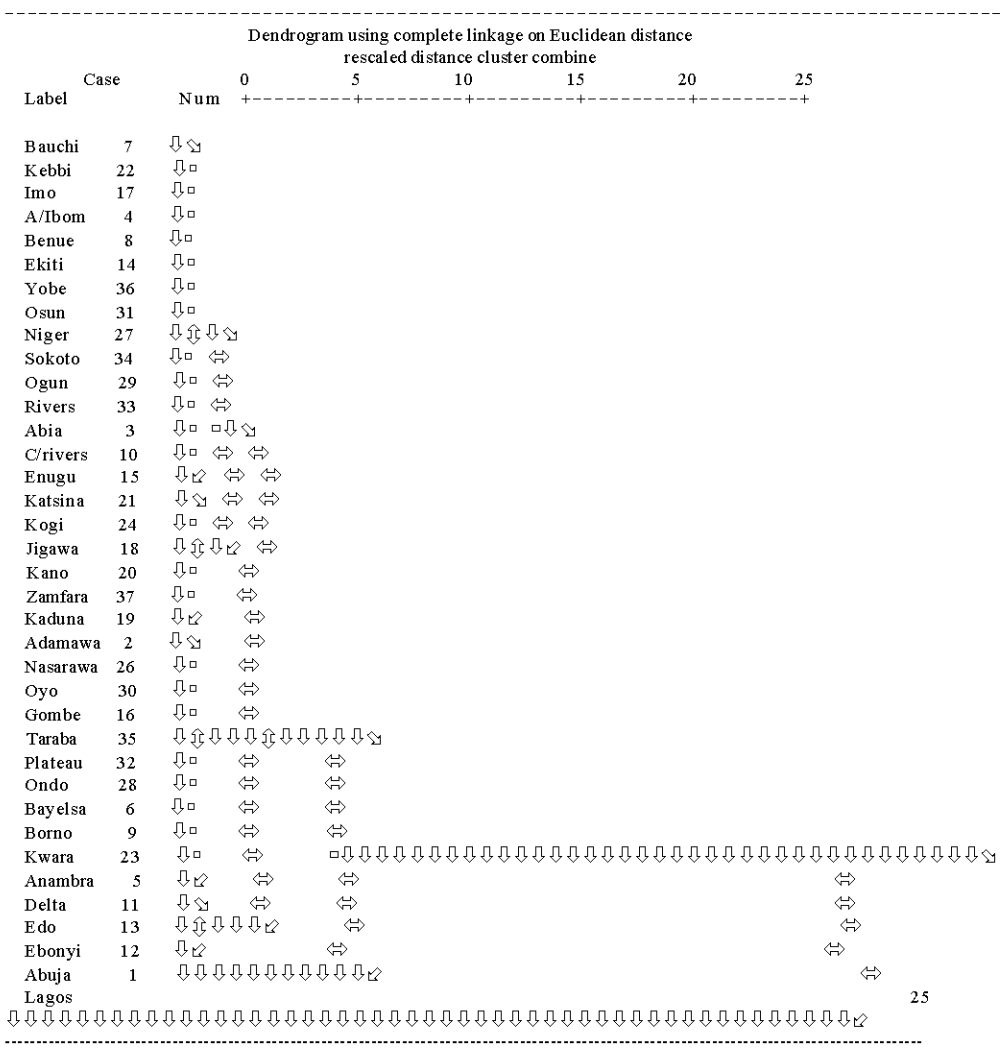

Fig. 1: Dendrogram of the complete linkage performed on the Euclidean distances

The states are then classified into higher and lower crime clusters using the averages of the standardized number of crimes which are shown in Table 2.

Table 2 shows the differences between the clusters. Cluster 1 and 2 contains the states with low criminality since the average of the standardized number of all crimes is negative except in Cluster 1 were they show slight tendency toward burglary. Furthermore, Cluster 1 has the lowest criminality since, it has the lowest average of the total standardized values of all crimes. Cluster 3 shows tendency toward burglary and store breakings. All the crimes are spread in Cluster 4 at different levels with the exception of theft/stealing, burglary and house breakings. Lagos state in Cluster 5 represents all crimes excluding robbery, vehicle theft and arson. Abuja in Cluster 6 has the highest criminality in Nigeria although, rate of burglary is very minimal. Robbery, forgery, arson, vehicle theft and false pretence and cheating appear to be higher in Abuja while theft/staling, store breakings and burglary in Lagos State. The number of clusters produced by performing the complete linkage on Manhattan distances are numerous and therefore could not serve the purpose of cluster analysis.

Ward Method: The result of the Ward Method performed on the Euclidean distances between standardized observations is shown in Fig. 2 and Table 3. Here six clusters are chosen for consideration and are shown in Table 4.

However, the first four clusters formed here are different from that of the complete linkage. Cluster 1 consists of the previous Cluster 1 plus the Kebbi, Bauchi and Imo States, Ekiti and Yobe States are added to Cluster 3 and C/Rivers State is added to Cluster 4. Similarly, Cluster 1 and 2 contains the states with low criminality since, the average of the standardized number

Table 1:Result of complete linkage on Euclidean distances

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
1,1 Katsina	2,12 Bauchi	2,3 Adamawa	6,5 Delta	4,2 Lagos	3,7 Abuja
1,1 Jigawa	1,10 Kebbi	3,4 Nasarawa	6,5 Edo		
1,1 Kano	5,9 Imo	4,11 Oyo	5,6 Ebonyi		
1,7 Kaduna	6,6 A/Ibom	2,3 Gombe			
1,10 Zamfara	3,4 Benue	2,3 Taraba			
3,8 Kogi	4,8 Ekiti	3,4 Plateau			
	2,12 Yobe	4,11 Ondo			
	4,11 Osun	6,5 Bayelsa			
	3,7 Niger	2,5 Borno			
	1,10 Sokoto	3,8 Kwara			
	4,2 Ogun	5,9 Anambra			
	6,6 Rivres				
	5,9 Abia				
	6,6 C/Rivers				
	5,9 Enugu				

Table 2: Average of the standardized Nigeria crime data set within the six clusters

Clusters	Robbery	Theft	V. theft	Burglary	H. break	S. break	FPC	Forgery	Arson	Total
1	-0.234	-0.371	-0.344	0.117	-0.566	-0.896	-0.574	-0.545	-0.394	-0.423
2	-0.216	-0.230	-0.215	-0.105	-0.288	-0.367	-0.289	-0.179	-0.254	-0.238
3	-0.088	-0.042	-0.208	0.399	0.046	0.383	-0.253	-0.145	0.028	0.013
4	0.358	-0.153	0.394	-0.689	-0.256	0.331	1.820	0.800	0.471	0.342
5	-0.529	5.816	-0.459	1.045	2.914	2.899	0.634	0.659	-0.554	1.380
6	5.287	0.651	5.332	-0.694	4.594	1.507	4.268	4.168	4.711	3.314

Table 3: Result of Ward Method on Euclidean distances

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
1,10 Kebbi	6,6 A/Ibom	2,3 Adamawa	6,5 Delta	4,2Lagos	3,7Abuja
2,12 Bauchi	3,4 Benue	3,4 Nasarawa	6,5 Edo		
5,9 Imo	5,9 Enugu	4,11 Oyo	6,6 C/Rivers		
1,1 Jigawa	3,7 Niger	4,8 Ekiti	5,6 Ebonyi		
1,1 Kano	1,10 Sokoto	2,12 Yobe			
1,10 Zamfara	4,11 Osun	4,11 Ondo			
1,7 Kaduna	4,2 Ogun	3,8 Kwara			
1,1 Katsina	6,6 Rivres	6,5 Bayelsa			
3,8 Kogi	5,9 Abia	2,5 Borno			
		2,3 Gombe			
		2,3 Taraba			
		3,4 Plateau			
		5,9 Anambra			

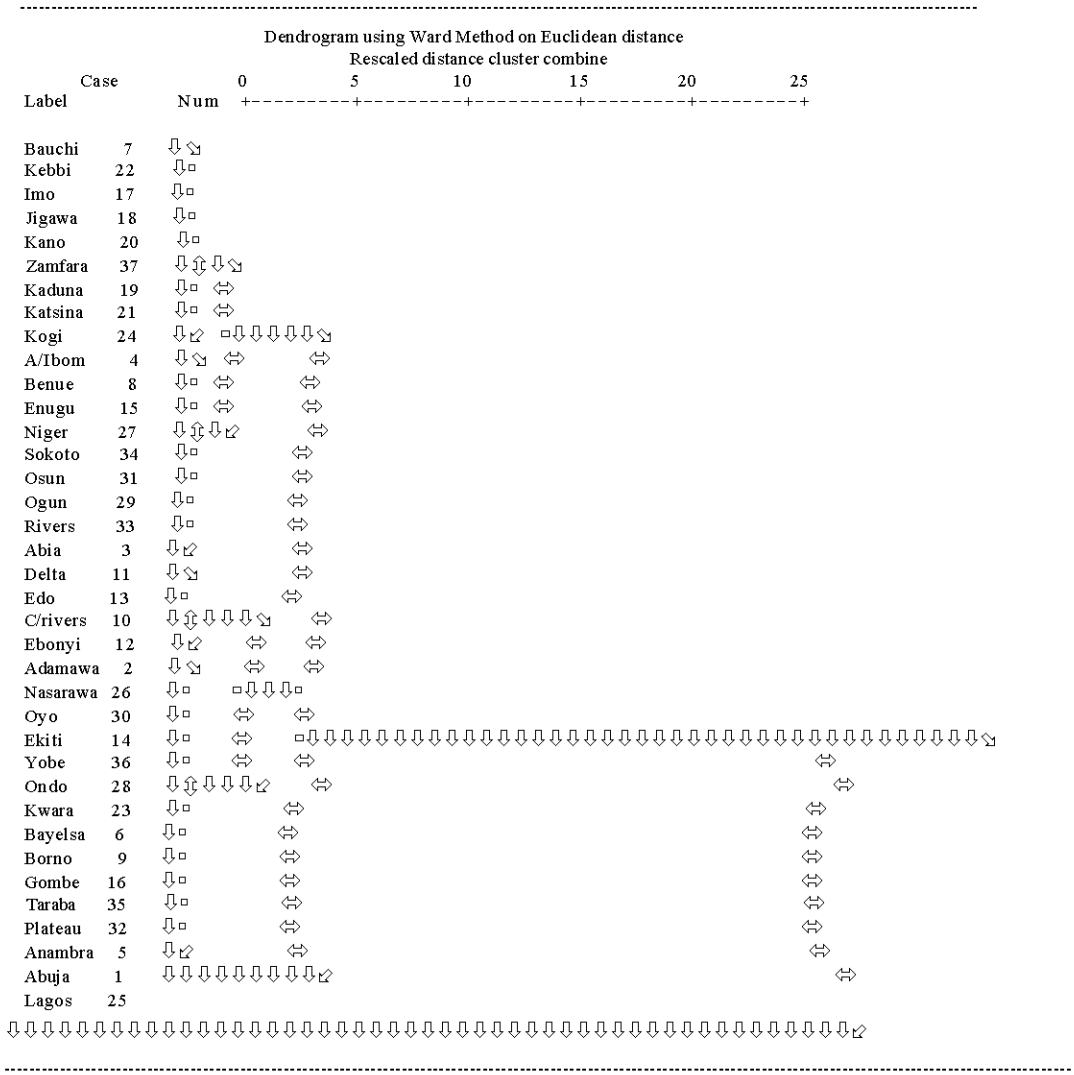

Fig. 2: Ward Method performed on the Euclidean distances between standardized observations

Table 4: Average of the standardized Nigeria Crime data set within the six clusters

Clusters	Robbery	Theft	V. theft	Burglary	H. break	S. break	FPC	Forgery	Arson	Total
1	-0.205	-0.339	-0.382	0.121	-0.501	-0.910	-0.543	-0.510	-0.274	-0.394
2	-0.339	-0.216	0.107	-0.353	-0.298	-0.260	-0.301	-0.031	-0.287	-0.220
3	-0.085	-0.058	-0.240	0.369	0.063	0.398	-0.289	-0.207	0.004	-0.005
4	0.310	-0.179	0.180	-0.753	-0.285	0.240	1.606	0.682	0.220	0.225
5	-0.529	5.816	-0.459	1.045	2.914	2.899	0.634	0.659	-0.554	1.380
6	5.287	0.651	5.332	-0.694	4.594	1.507	4.268	4.168	4.711	3.314

of almost the crimes is negative although, Cluster 1 shows slight tendency toward burglary and Cluster 2 shows slight tendency toward vehicle theft. Again, Cluster 1 has the lowest criminality since it has the lowest average of the total standardized values of all crimes. The interpretations for the four remaining clusters remain the same with that of complete linkage.

The result of the Ward Method performed on the Manhattan distance between standardized observations is shown on Fig. 3 and Table 5. Six clusters are chosen for consideration and are shown on Table 6. The first two clusters of lower crime states of Ward Method on Euclidean distances are now partition in to three by performing Ward Method on Manhattan distances.

Cluster 1 consisting of all the states in Zone 1 command and most of the North Western states has the lowest crime rates, although with slight tendency toward burglary. Similarly, Cluster 2 consisting of two states from

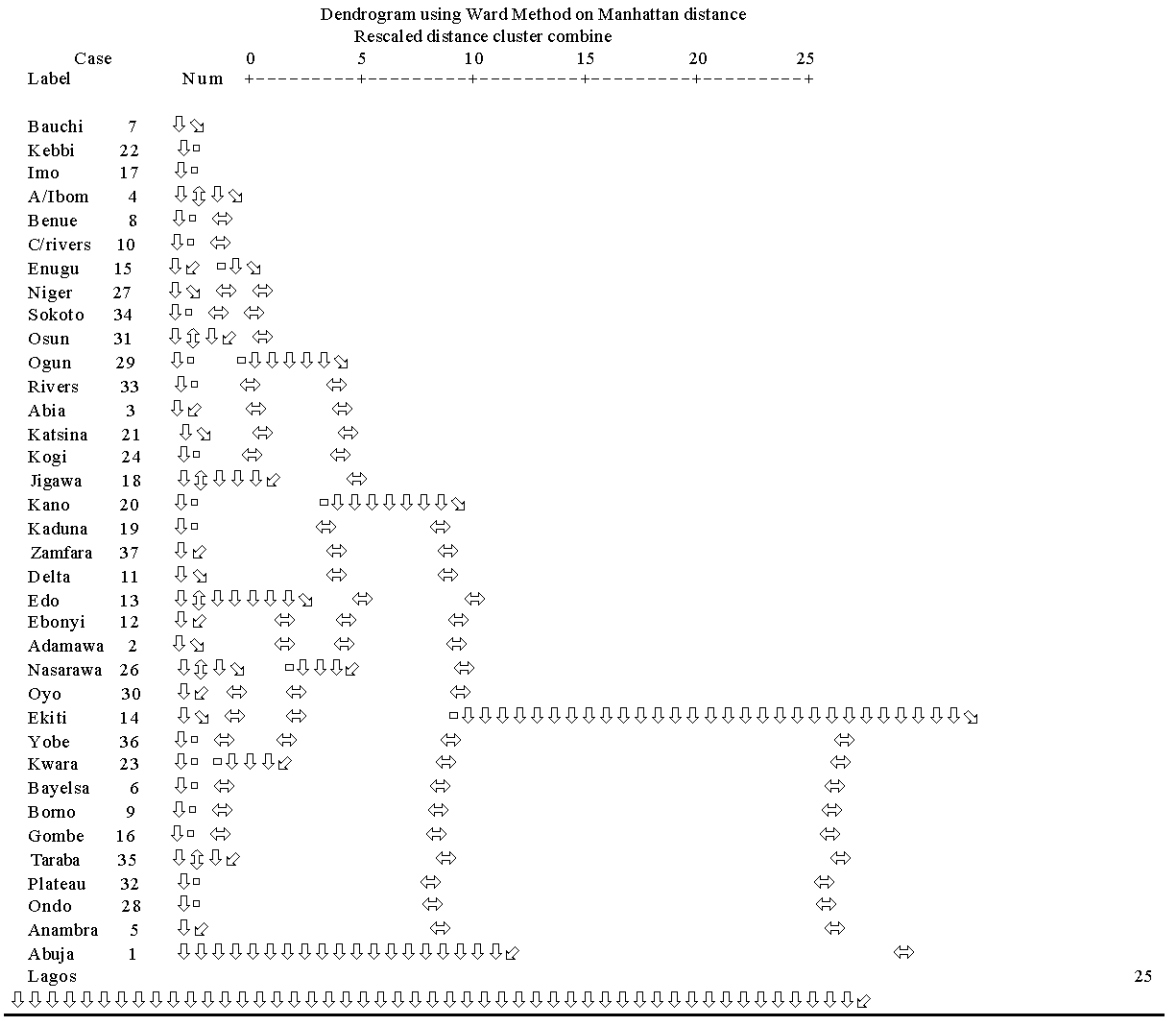

Fig. 3: Ward Method performed on the Manhattan distance between standardized observation

Table 5: Result of Ward Method on Manhattan distances

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
1,1 Katsina	2,12 Bauchi	3,7 Niger	2,3 Adamawa	4,8 Ekiti	6,5 Delta	4,2 Lagos	3,7 Abuja
3,8 Kogi	1,10 Kebbi	1,10 Sokoto	3,4 Nasarawa	2,12 Yobe	6,5 Edo		
1,1 Jigawa	5,9 Imo	4,11 Osun	4,11 Oyo	4,11 Ondo	5,6 Ebonyi		
1,1 Kano	6,6 A/Ibom	4,2 Ogun		3,8 Kwara			
1,7 Kaduna	3,4 Benue	6,6 Rivres		6,5 Bayelsa			
1,10 Zamfara	6,6 C/Rivers	5,9 Abia		2,5 Borno			
	5,9 Enugu			2,3 Gombe			
				2,3 Taraba			
				3,4 Plateau			
				5,9 Anambra			

Table 6: Average of the standardized Nigeria Crime data set within the eight clusters

Cluters	Robbery	Theft	V. theft	Burglary	H. break	S. break	FPC	Forgery	Arson	Total
1	-0.234	-0.371	-0.344	0.117	-0.566	-0.896	-0.574	-0.545	-0.394	-0.423
2	-0.102	-0.261	-0.374	-0.412	-0.392	-0.632	-0.060	-0.252	-0.114	-0.289
3	-0.434	-0.199	0.291	-0.142	-0.237	-0.127	-0.461	0.083	-0.404	-0.181
4	-0.438	0.056	-0.223	-0.525	0.336	-0.108	-0.213	-0.267	0.489	-0.099
5	0.020	-0.092	-0.245	0.637	-0.019	0.550	-0.312	-0.189	-0.141	0.023
6	0.358	-0.153	0.394	-0.689	-0.256	0.331	1.820	0.800	0.471	0.342
7	-0.529	5.816	-0.459	1.045	2.914	2.899	0.634	0.659	-0.554	1.380
8	5.287	0.651	5.332	-0.694	4.594	1.507	4.268	4.168	4.711	3.314

Zone 6 command and two states each from South-South and South Eastern states has lower crime rates. By comparing Clusters 4 and 6 of Table 1 and 3, respectively and the standardized values of all the crimes against C/Rivers states although, the state is classified as lower crimes state, it shows tendency towards FPC and forgery. Again, Cluster 3 consisting of two South Western states has low crime rates with tendency towards vehicle theft.

This method has also partition Cluster 3 of Table 2 in to two more clusters to identify the states that influence the status of the cluster. Here while Cluster 3 consisting of Adamawa, Nasarawa and Oyo states shows tendency towards arson and house breakings, Cluster 5 has retained the spread of the usual burglary and store breakings. The interpretations of the last three clusters remain the same as that of the complete linkage.

CONCLUSION

This study has identified the usefulness of applying different agglomerative algorithms on different distance measures simultaneously in producing better classifications of areas into higher and lower property crime concentrations across the 36 states in Nigeria. Although, selection of suitable cluster classification from different methods is subjective in this study however, the cluster classification by performing Ward Method on Manhattan distance is viable.

Nineteen states appeared to have low property criminality in Nigeria which can be categorized into three levels:

- Katsina, Kogi, Jigawa Kano, Kaduna and Zamfara States have the lowest average property crime rates however, they show tendency towards burglary
- Bauchi, Kebbi, Imo, Akwa Ibom, Benue, Cross Rivers and Enugu States have lower crime rates
- Niger, Sokoto, Osun, Ogun, rivers and Abia States have low crime rates with tendency towards vehicle theft
- All the states in North-West region and under Zone 1 command have low criminality while only Bauchi state from North-East region has low criminality
- The following states from their respective regions have low crime rates: Benue, Kogi and Niger States from North-Central, Osun and Ogun States from South-West, Imo, Enugu and Abia States from South-East and Akwa Ibom, Cross Rivers and Rivers States from South-South
- All the states under Zone 6 command have low criminality except Ebonyi State

Abuja has the highest criminality and is followed by Lagos State. Robbery, vehicle theft, house breaking, false pretence and cheating, forgery and arson are largely spread in Abuja while, theft, burglary and store breakings are dominant in Lagos State.

REFERENCES

Cahill, M.E., 2004. Geographies of urban crime: An intraurban study of crime in Nashville, TN; Portland, OR and Tucson, AZ. Ph.D. Thesis, Department of Geography and Regional Development, University of Arizona, Tucson, Arizona.

Fajemirokun, F., O. Adewale, T. Idowu, A. Oyewusi and B. Maiyegun, 2006. A GIS approach to crime mapping and management in Nigeria: A case study of Victoria Island Lagos. Proceedings of the Conference on Shaping the Change and XXIII FIG Congress, October 8-13, 2006, Munich, Germany, pp: 1-17.

Hardle, W. and H. Zdenek, 2007. Multivariate Statistics: Exercises and Solutions. Springer-Verlag, New York.

Hardle, W. and L. Simar, 2003. Applied Multivariate Statistical Analysis. 2nd Edn., Springer-Verlag, New York.

Lance, G.N. and W.T. Williams, 1967. A general theory of classificatory sorting strategies. 1. Hierarchical Syst. Compt. J., 9: 373-380.

Lasch, P., W. Haensch, D. Naumann and M. Diem, 2004. Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. Biochim. Biophys. Acta, 1688: 176-186.

Maydeu-Olivares, A., 1996. Classification of prison inmates based on hierarchical cluster analysis. Psicothema, 8: 709-715.

NPF, 2009. The annual report of the Nigeria police force 2009. 'F' Department, Force Headquarters, Abuja.

Radoff, D., 1993. GIS responds to emergency management. Int. J. GIS, 34: 209-210.

Rencher, A.C., 2002. Methods of Multivariate Analysis. 2nd Edn., John Wiley and Sons, New York.

Tsiamtsiouri, A. and J. Panaretos, 1999. Some statistical analysis of greek crime data. Proceedings of the International Conference Envirometrics and Statistics in the Earth and Space Sciences, August, 1999, Athens, Greece.

Ward, J.H., 1963. Hierarchical grouping methods to optimize an objective function. J. Am. Stat. Assoc., 58: 236-244.