Analysis of Crime Data using Principal Component Analysis: A case study of Katsina State

Shehu U. Gulumbe[1], H.G. Dikko[2], and Yusuf Bello[3]

This paper analyses Katsina State crime data which consists of the averages of eight major crimes reported to the police for the period 2006 – 2008. The crimes consist of robbery, auto theft, house and store breakings, theft/stealing, grievous hurt and wounding, murder, rape, and assault. Correlation analysis and principal component analysis (PCA) were employed to explain the correlation between the crimes and to determine the distribution of the crimes over the local government areas of the state. The result has shown a significant correlation between robbery, theft and vehicle theft. While MSW local government area has the lowest crime rate, KTN local government area has the overall crime rate in the state. Robbery is more prevalent in DMS local government area, rape in JBA local government area, and grievous hurt and wounding in DDM local government area. The PCA has suggested retaining four components that explain about 78.94 percent of the total variability of the data set.

JEL Classification: C2,C5,C13

Keywords: Crime data; Correlation analysis; PCA

1.0 Introduction

Crime is one of the continuous problems that bedevil the existence of mankind. Since forth early days, crime had been a disturbing threat to his personality, property and lawful authority (Louis *et al.*, 1981). Today, in the modern complex world, the situation is most highly disturbing. Crime started in the primitive days as a simple and less organised issue, and ended today as very complex and organised. Therefore, the existence of crime and its problems have spanned the history of mankind.

Nigeria has one of the alarming crime rates in the world (Uche, 2008 and Financial, 2011). Cases of armed robbery attacks, pickpockets, shoplifting and 419 have increased due to increased poverty among population (Lagos, undated). In the year 2011, armed robbers killed at least 12 people and

[1] Department of Mathematics, Usmanu Danfodiyo University, Sokoto.

[2] Department of Mathematics, Ahmadu Bello University, Zaria.

[3] Department of Mathematics and Statistics, Hassan Usman Katsina Polytechnic, Katsina.
Email: byusuf10@yahoo.com, Mobile: 08082586690

possibly more in attacks on a bank and police station in North-Eastern Nigeria (Nossiter, 2011). However, Maritz (2010) has considered that image as merely exaggeration. He added that, as is the case with the rest of the world, Nigeria's metropolitan areas have more problems with crime than the rural areas. Most crimes are however, purely as a result of poverty.

Despite the fact that, crime is inevitable in a society Durkheim (1933), various controlling and preventive measures had been taken, and are still being taken to reduce the menace. However, crime control and prevention is still bedevilled by numerous complex problems. When an opportunity for crime is blocked, an offender has several alternative types of displacement (Gabor, 1978). However, the introduction of modern scientific and technical methods in crime prevention and control has proved to be effective. The application of multivariate statistics has made some contributions to many criminological explanations (Kpedekpo and Arya, 1981 and Printcom, 2003).

This paper explores the use of correlation analysis and PCA for effective crime control and prevention. PCA offers a tool for reducing the dimensionality of a very large data set and in determining the areas with overall crime rate. These if properly implemented, will successively solve many of the complex criminal problems that have bedevilled the country in general and Katsina State in particular.

2.0 Methodology

The crime data for the period 2006 – 2008 for the 36 Divisional Police Headquarters (DPHs) or Local Government Areas (LGAs) of Katsina State was collected officially from the record of Statistics (F) Department of the Nigeria Police Force, Katsina State Command. For easy statistical analysis and interpretation, the 36 LGAs were categorized according to the three existing Area Commands (ACs): Katsina, Funtua and Daura Area Commands. Each LGA within the three ACs would be identified by 1, 2, and 3 respectively. The ACs and the LGAs within their category are as follows:

- *Katsina Area Command:* KTN, BAT, KTA, RMY, JBA, BTR, CRC, KUF, DTM, SFN and DMS LGAs.
- *Funtua Area Command*: FTA, BKR, FSK, DDM, SBW, DJA, KFR, MLF, KKR, MSW and MTZ LGAs.

- *Daura Area Command*: DRA, DTS, MDA, ZNG, SDM, BRE, ING, MSH, MAN, BDW, KSD and KNK LGAs.

The data consists of eight major crimes reported to the police for the period 2006 – 2008. The crime classifications are: Crimes against properties which include robbery, auto theft, house and store breakings and theft/stealing, and crimes against persons which include grievous hurt and wounding (G.H.W.), murder, rape, and assault.

Frequencies of crimes for each category were averaged over the three years in the study period to control for anomalous years when there may have been an unexplained spike or fall in crime levels prior to the statistical analysis. The value for each crime was converted to crime rate per 100,000 populations of the LGA which was calculated as (Kpedekpo and Arya, 1981):

$$Crime \;\; rate = \frac{number \;\, of \;\, crime \; committed}{population \;\, of \;\, the \;\, LGA} * 100\ 000 \tag{1.1}$$

3.0 Principal Component Analysis

Let X be a vector of p random variables, the main idea of the PC transformation is to look for a few ($< p$) derived variables that preserved most of the information given by the variance of the p random variables (Jolliffe, 2002). Let the random vector $X' = [X_1, X_2, \dots, X_p]$have the covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Consider the linear combinations

$Y_j = \alpha_j' X = \alpha_{j1}X_1 + \alpha_{j2}X_2 + ... + \alpha_{jp}X_p = \sum_{k=1}^{p} \alpha_{jk}X_k, \qquad j = 1,2,..., p$ of the element of X, where α_j is a vector of p components $\alpha_{j1}, \alpha_{j2}, \dots, \alpha_{jp}$.

Then

$$Var(Y_j) = \alpha_j' \Sigma \alpha_j \qquad j = 1,2,..., p \tag{1.2}$$

$$Cov(Y_j, Y_k) = \alpha_j' \Sigma \alpha_k \qquad j,k = 1,2,..., p \tag{1.3}$$

The PCs are those *uncorrelated* linear combinations $Y_1, Y_2, \dots, Y_p$ whose variances in (1.2) are as large as possible (Richard and Dean, 2001). In finding the PCs we concentrate on the variances. The first step is to look for a linear combination $\alpha_1'\boldsymbol{X}$ with maximum variance, so that

$$\alpha_1'X = \alpha_{11}X_1 + \alpha_{12}X_2 + \ldots. + \alpha_{1p}X_p = \sum_{k=1}^{p} \alpha_{1k}X_k$$

Next, look for a linear combination $\alpha_2'\boldsymbol{X}$ uncorrelated with $\alpha_1'\boldsymbol{X}$ having maximum variance, and so on, so that at the k^{th} stage a linear combination $\alpha_k'\boldsymbol{X}$ is found that has maximum variance subject to being uncorrelated with $\alpha_1'\boldsymbol{X}, \alpha_2'\boldsymbol{X}, \ldots, \alpha_{k-1}'\boldsymbol{X}$. The k^{th} derived variable $\alpha_k'\boldsymbol{X}$ is the k^{th} PC. Up to p PCs could be found, but we have to stop after the q^{th} stage ($q \le p$) when most of the variation in X have been accounted for by q PCs.

- The variance of a PC is equal to the eigenvalue corresponding to that PC,

$$Var(Y_j) = \alpha_j'\Sigma\alpha_j = \lambda_j \qquad j = 1,2,\ldots, p$$

- The total variance in a data set is equal to the total variance of PCs

$$\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \sum_{j=1}^{p} Var(X_j) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{j=1}^{p} Var(Y_j)$$

The data was standardized for the variables to be of similar scale using a common standardization method of transforming all the data to have zero mean and unit standard deviation. For a random vector $X' = [X_1, X_2, \ldots, X_p]$ the corresponding standardized variables are

$$Z = \left[Z_j = \frac{(X_j - \boldsymbol{\mu}_j)}{\sqrt{\boldsymbol{\sigma}_{jj}}} \right] \qquad j = 1,2,\ \ldots,\ p$$

In matrix notation,

$$Z = \left(\boldsymbol{V}^{1/2}\right)^{-1} (\boldsymbol{X} - \boldsymbol{\mu})$$

where $V^{1/2}$ is the diagonal standard deviation matrix. Thus $E(Z) = 0$ and $Cov(Z) = \rho$.

The PCs of Z can be obtained from eigenvectors of the correlation matrix ρ of X. All our previous properties for X are applied for the Z, so that the notation

Y_j refers to the j^{th} PC and (λ_j, α_j) refers to the eigenvalue – eigenvector pair. However, the quantities derived from Σ are not the same from those derived from ρ (Richard and Dean, 2001).

The j^{th} PC of the standardized variables $Z' = [z_1, z_2, \dots, z_p]$ with $cov\ (Z) = \rho$, is given by

$$Y_j = \alpha_j' Z = \alpha_j' \left(V^{1/2}\right)^{-1} (X - \mu),$$

so that

$$\sum_{j=1}^{p} Var(Y_j) = \sum_{j=1}^{p} Var(Z_j) = p \qquad j = 1,2,\dots, p$$

In this case, $(\lambda_1, \alpha_1), (\lambda_2, \alpha_2), \dots (\lambda_p, \alpha_p)$ are the eigenvalue- eigenvector pairs for ρ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Interpretation of the Principal Components:

The *loading* or the eigenvector $\alpha_j = \alpha_1, \alpha_2, \dots, \alpha_p$, is the measure of the importance of a measured variable for a given PC. When all elements of α_1 are positive, the first component is a weighted average of the variables and is sometimes referred to as measure of *overall crime rate*. Likewise, the positive and negative coefficients in subsequent components may be regarded as *type of crime* components (Rencher, 2002 and Printcom, 2003). The plot of the first two or three loadings against each other enhances visual interpretation (Soren, 2006).

The *score* is a measure of the importance of a PC for an observation. The new PC observations Y_{ij} are obtained simply by substituting the original variables X_{ij} into the set of the first q PCs. This gives

$$Y_{ij} = \alpha_{j1}' X_{i1} + \alpha_{j2}' X_{i2} + \dots + \alpha_{jp}' X_{ip} \qquad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, p$$

The plot of the first two or three PCs against each other enhances visual interpretation (Soren, 2006).

The proportion of Variance:

The proportion of variance tells us the PC that best explained the original variables. A measure of how well the first q PCs of Z explain the variation is given by

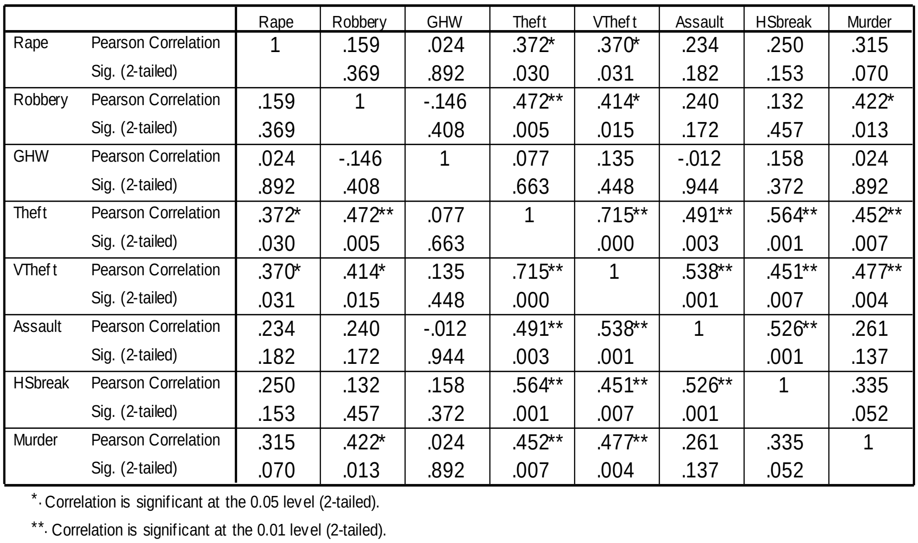
$$\psi_q = \frac{\sum_{j=1}^{q}\lambda_j}{P} = \frac{\sum_{j=1}^{q}Var(Z_j)}{P}$$

A cumulative proportion of explained variance is a useful criterion for determining the number of components to be retained in the analysis. A Scree plot provides a good graphical representation of the ability of the PCs to explain the variation in the data (Cattell, 1966).

4.0 Analysis and Results

The correlation matrix in Table 1 has displayed different levels of correlation between the crimes. There is no significant correlation in between Crimes against Persons which means that none of the variables can be used to predict (explain) one another. However, the correlations in between Crimes against Property is at least moderate except between house breaking and robbery, and therefore each crime can be used to predict (explain) one another.

Table 1: Correlation of Crime Types (per 10,000 population) in Katsina State



		Rape	Robbery	GHW	Theft	VTheft	Assault	HSbreak	Murder
Rape	Pearson Correlation	1	.159	.024	.372*	.370*	.234	.250	.315
	Sig. (2-tailed)		.369	.892	.030	.031	.182	.153	.070
Robbery	Pearson Correlation	.159	1	-.146	.472**	.414*	.240	.132	.422*
	Sig. (2-tailed)	.369		.408	.005	.015	.172	.457	.013
GHW	Pearson Correlation	.024	-.146	1	.077	.135	-.012	.158	.024
	Sig. (2-tailed)	.892	.408		.663	.448	.944	.372	.892
Theft	Pearson Correlation	.372*	.472**	.077	1	.715**	.491**	.564**	.452**
	Sig. (2-tailed)	.030	.005	.663		.000	.003	.001	.007
VTheft	Pearson Correlation	.370*	.414*	.135	.715**	1	.538**	.451**	.477**
	Sig. (2-tailed)	.031	.015	.448	.000		.001	.007	.004
Assault	Pearson Correlation	.234	.240	-.012	.491**	.538**	1	.526**	.261
	Sig. (2-tailed)	.182	.172	.944	.003	.001		.001	.137
HSbreak	Pearson Correlation	.250	.132	.158	.564**	.451**	.526**	1	.335
	Sig. (2-tailed)	.153	.457	.372	.001	.007	.001		.052
Murder	Pearson Correlation	.315	.422*	.024	.452**	.477**	.261	.335	1
	Sig. (2-tailed)	.070	.013	.892	.007	.004	.137	.052	

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Source: Derived from Statistics (F) Department of the Nigeria Police Force, Katsina State Command

Murder has shown significant correlations between robbery, theft and vehicle theft. This means that the high rate of murder in the state is associated to property crime.

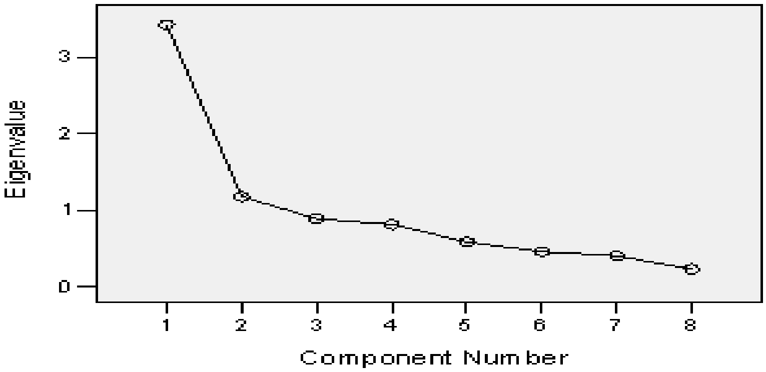


Fig. 1: Scree Plot

The eigenvalues and the cumulative proportions of the explained variance are displayed in Table 2. Considering the eigenvalue-one criterion and the Scree plot in figure 1, it would be reasonable to retain the first two PCs. A commonly accepted rule says that it suffices to keep only PCs with eigenvalues larger than 1. However, the third and the forth eigenvalues $\lambda_3 = 0.891$and $\lambda_4 = 0.815$ are approximately close to 1, so that the first 4 PCs can be retain to explain up to 78.938 percent of the total variability.

Table 2: Eigenvalues

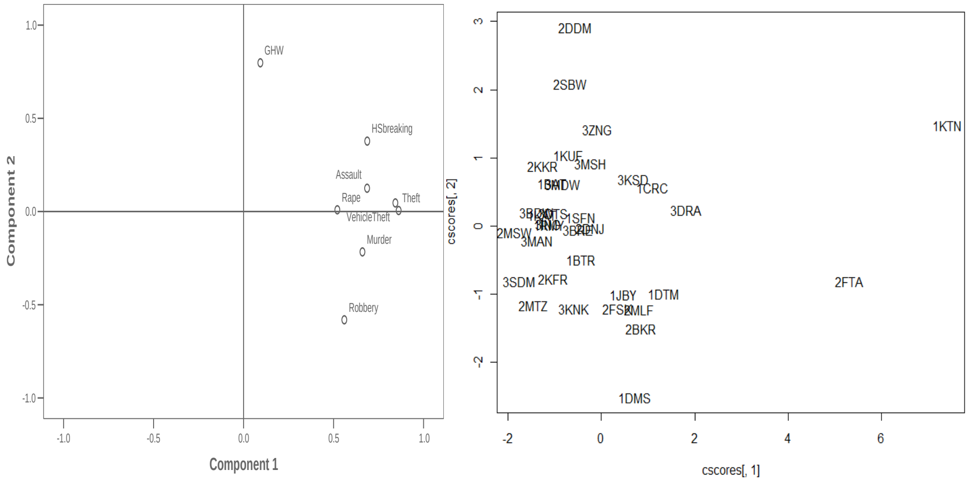
Component	Eigenvalues	Proportion	Cumulative
1	3.427	42.842	42.842
2	1.181	14.769	57.611
3	0.891	11.136	68.746
4	0.815	10.191	78.938
5	0.588	7.349	86.287
6	0.462	5.777	92.063
7	0.405	5.056	97.120
8	0.230	2.888	100.000

From Table 3, the first PC combines the number of all the crimes with approximately positive constant (0.05 – 0.47) weight, and is interpreted as the overall measure of crime. From figure 2b, KTN Local Government area has the overall crime rate, while MSW and SDM Local Government areas have the lowest crime rate. The second PC on figure 2a has classified the crimes into groups: (1) the concentrated crimes (consisting of felonies and misdemeanours offences): G.H.W., theft, assault, rape, house and Store breakings and vehicle theft, and (2) felonies: murder and robbery. The third component is difficult to be interpreted.

Table 3: Eigenvectors

	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8
Rape	0.281	0.008	-0.322	-0.853	0.190	0.043	0.224	-0.031
Robbery	0.302	-0.534	-0.211	0.392	0.292	0.191	0.522	-0.174
G.H.W.	0.051	0.734	-0.506	0.311	0.174	-0.101	0.247	0.070
Theft	0.465	0.001	0.015	0.093	0.190	0.400	-0.356	0.673
V.Theft	0.456	0.043	-0.048	0.091	0.283	-0.238	-0.568	-0.567
Assault	0.371	0.115	0.565	-0.006	0.136	-0.584	0.347	0.227
H/Sbrea	0.371	0.348	0.342	-0.003	-0.425	0.527	0.210	-0.349
Murder	0.357	-0.199	-0.397	0.069	-0.730	-0.341	-0.044	0.139

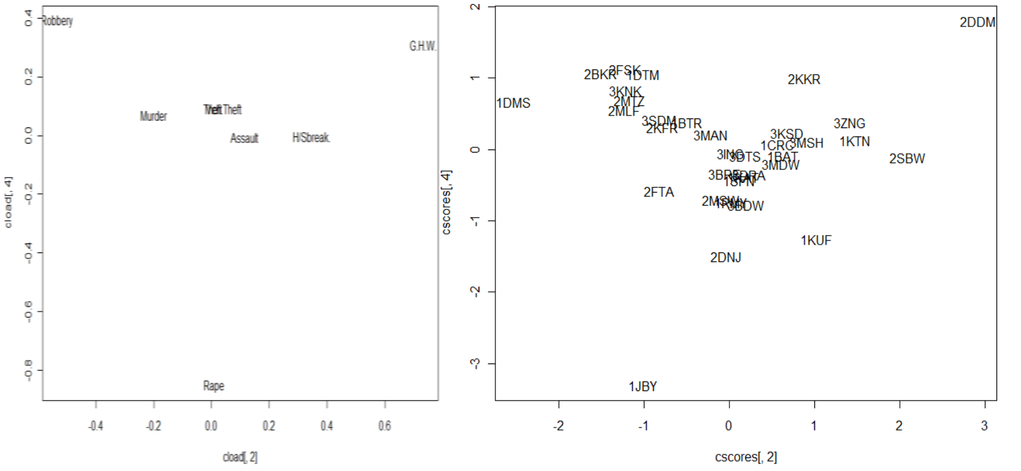
From Figure 3a, rape is an outlier and is located at the lowest side, and G.H.W. is located at the right part. Therefore, from Figure 3b, the Local Government areas at the lower side show tendency towards rape, where JBA Local Government area has the highest prevalence. SBW, ZNG and KTN show tendency towards G.H.W., where DDM Local Government area has the highest prevalence.



(a) **(b)**

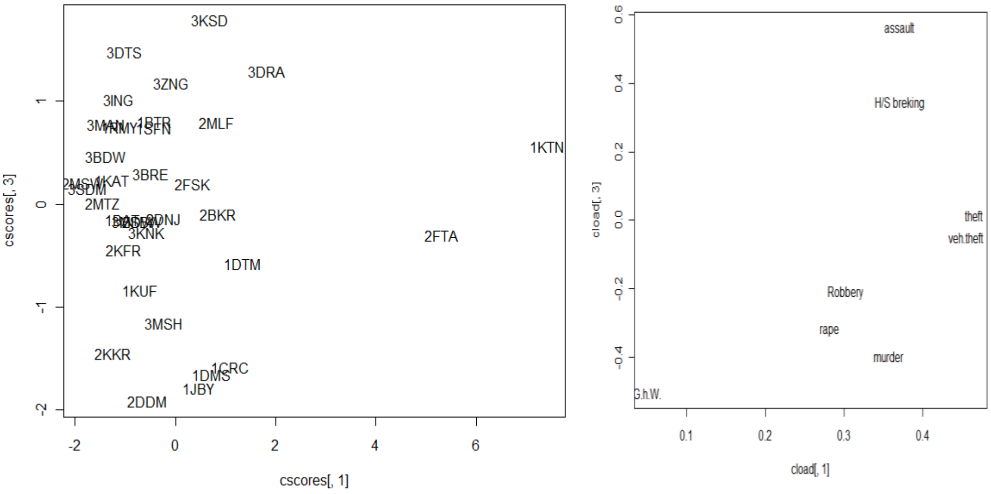
Fig. 2(a): Loading plot for the first and second PC **(b)** Score plot for the first and second PC

Assault and house and store breakings are located at the upper part of figure 4a, and therefore the corresponding Local Government areas at the upper part of figure 4b including KSD, DTS, DRA, ZNG and ING Local Government areas show tendency toward assault and to some extent house and store breakings. All these Local Government areas are located in AC 3.



(a) (b)

Fig. 3(a): Loading plot for the second and third PC **(b)** Score plot for the second and third PC



(a) (b)

Fig. 4(a): Loading plot for the first and third PC **(b)** Score plot for the first and third PC

5.0 Conclusion

The following are the conclusions deduced from the paper. There are no significant correlations among Crimes against Persons which means that none of the variables can be used to predict (explain) one another. The correlations among Crimes against Property are at least moderate. However, there are significant correlations among robbery, theft and vehicle theft. The Local Government with the highest crime rate KTN, while MSW and SDM Local Governments have the lowest crime rate in the state.

The second component suggested that DDM Local Government area has the highest G.H.W. cases, while robbery and murder are the popular crimes in the AC2 and the Local Government areas located at the southern part of AC1. The forth component suggested that rape is very predominant in JBA Local Government area.

Four PCs that explains about 78.94 per cent of the total variability of the data set are suggested to be retained. The second component has classified the crimes into two, namely, (1) concentrated offences: theft, G.H.W., vehicle theft, assault, house and Store breakings and vehicle theft, (2) felonies: murder and robbery. Base on this, the component has geographically divided the state between the north and south in relation to the crime classifications. The southern parts of AC1 and 3 toward the southerly AC2 Local Government areas contain more murder and robbery, while the northern part has the prevalence of the concentrated crimes. Thus, identifying the distribution of crimes in Katsina State allows the investors to measure the level of risk and to plan preventive measures for safeguarding their investments.

Reference:

Cattell, R.B. (1966). The Scree test for the number of factors. *Multivariate Behavioral Research, 1,*245-276.

Durkheim, E. (1933). The division of Labour in Society. Macmillan, New York.

Financial (2011). Nigeria crime. *Financial Times*, 7/11/ 2011.

Gabor, T. (1978). Crime displacement: the literature and strategies for its investigation. *Crime and Justice,* 6:100 - 7.

Jolliffe, I.T. (2002). Principal Component Analysis. 2nd edn, Springer-Verlag, New York.

Kpedekpo, G.M.C. and Arya, P. L. (1981). Social and Economic Statistics for Africa. George Allen and Unwin, London.

Lagos (undated). Crime Rate in Nigeria – Free Tips to Deal with Armed Robbery Attacks, Pickpockets, Shoplifting, And 419. http://lagos-nigeria-real-estate-advisor.com/crime-rate.html

Louis, S., Cookie, W. S., Louis, A. Z. and Sheldon, R. E. (1981). Human Response to Social Problems. The Dorsey Press, Illinois.

Maritz, J. (2010). Honest answers to your questions about investing in Nigeria: Will I have to fear for my safety in Nigeria? June, 14, 2010.http://www.tradeinvest nigeria. com/news/623915.htm

Nossiter, A. (2011). Robbers kill at least 12 in Nigeria. August 25, 2011, http://www.nytimes.com/2011/08/26/world/africa/26nigeria.html?_r=1

Printcom (2003) http://support.sas.com/onlinedoc/912/getDoc/common.hlp/ images/copyrite.htm.

Rencher, A.C. (2002) Methods of Multivariate Analysis. 2nd edn, John Wiley & Son, New York.

Richard, A.J. and Dean, W.W. (2001). Applied Multivariate Statistical Analysis. 3rd edn, Prentice-Hall, New Delhi.

Risk (undated). Nigeria Risk Assessment: Crime. http://www.professionaltravelguide.com/Destination/Nigeria/Safety/Risk-Assessment/crime/

Soren, H. (2006). Example of multivariate analysis in R – Principal component analysis (PCA). httpgene://tics.agrsci.dk/statistics/courses/Rcourse-Djf2006 / day3/ PCA-notes.

Uche, O. (2008). Nigeria Prison Robbed by Criminals. http://www.whichwayNigeria.net /Nigerian-prison-robbed-criminals/