PRINCIPAL COMPONENT ANALYSIS OF PAKISTAN CRIME DATA

Mehboob Ali,
Government Post Graduate College, Jauharabad District Khushab - Pakistan.
Email: Statistician_955@yahoo.com
Humera Razzak,
Department of Statistics, Government College University Faisalabad - Pakistan.
Email: ali.arif89@yahoo.com

***ABSTRACT*:** *This study analysis eight major crime data in Pakistan reported to the police for the period of January, 2005 - December, 2012. Principal component analysis (PCA) and correlation analysis is used to interpret crime data. Moderate correlations are found between crimes against property which means the prediction of each crime can be made by other variables. Same is the case with crime against a person where moderate correlation is found between crimes, except kidnaping and murder, which are strongly correlated and can be used to predict each other which means, high rate of murder in a state is associated to kidnappings. 89.87% of the total variability in the data set has been explained by retaining first three principal components (PCs). The overall crime rate of theft is very low and robbery has the highest crime rate. Second PC shows murder, burglary, kidnapping, other theft and dacoity has over all more crime prevalence.*

Keywords: *Crime date, Pakistan Bureau of Statistics, Principal component analysis.*

INTRODUCTION

The high crime rate in any nation is a major source of insecurity and fear to the public welfare and morals of its citizens. The Crime rate has a paramount importance to make judgment about the quality of life and welfare of any nation. Crime rate directly influences various decisions like purchasing property, relocating for jobs and starting new business, etc. and brings the economy level to its lowest ebb by making citizens lives a living hell. Crime can be described in different types as civil crime or social crime, etc. Crimes are conducted in every nation in the world either at a micro level or at a larger scale.

Central America experienced a marketable increase in the crime rate of murders in relation with high levels of organized crime since 2007. This has resulted in one of the highest sub-regional homicide rates in the world (26.5 per 100000 population) [1]. According to the survey conducted by United Nations office on drugs and crime in 2011 United States of America has the highest rate (12408899/100000) of crime followed by Germany 2112843 and France 1172547 [2]. According to the British crime survey, an increase of 190% mobile phone theft between 1995 and 2000 is reported, representing 28% of all robberies in 2000/2001 compared to 8% in 1998/1999 [3]. Central and Eastern Europe faced increase in drug and property offenses between 1990 and 2000. In addition to the figures mentioned above crime rate was highest during 1990s [4]. One of the highest crime rates in the world are found in Nigeria [5]. Incidence of crime by property and by self are positively correlated to the increase in poverty among population. South Africa also has a high prevalence of murder and violent crime rate (30.9) per 100,000 population [6].

Along with other major security issues, high crime rate is one of the particular problem which is endemic in Pakistan. The high crime rate in Islamic Republic of Pakistan is becoming alarming day by day involving various internal and external factors as well. The total number of reported crimes including dacoity, robbery, burglary, cattle theft, murder/attempted murder, kidnapping has gone up by about 63 percent during the period 1996-2007 [7]. Eight of the ten districts reporting highest number of crimes were in Punjab and two in Khyber Pakhtunkhwa (KPK). Districts with highest crime reporting includes Lahore (5102), Faisalabad (2294) and Peshawar (1665). Percentages of crimes pertaining to property, robbery and dacoity, and criminal trespass went up increased by 10% and 11%, respectively [8]. Mistrust between public and police serves an additional factor for many incidences of crimes to be remaining unreported.

Pakistan Bureau of Statistics (PBS) is a central statistical office of the government of Pakistan, supply of statistical information for analysis, compilation and collection of data is one of its main functions. PBS has published statistical data of incidence of personal and household victimization for 2005 to 2012, which serves as a tool for observing perception of the general public toward safety and security in the country. Data obtained from PBS can be categorized into two groups namely crime by property: dacoity, robbery, burglary, cattle theft, other theft and crime by a person: murder, attempted murder, kidnapping/abduction.

Principal component analysis has been used in other research on different types of criminal activities to derive crime components [9,10,11]. PCA finds an underlying dimension that explains the correlation among a set of variables [12]. It is a method that projects a dataset to a new coordinate system by determining the eigenvectors and eigen values of a matrix [13]. Effective crime control and prevention using correlation analysis and PCA has been explored in this paper. 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 major crimes related issues in the country [14].

METHODOLOGY

Data collection

Monthly data on crime based on police report was obtained from Pakistan Bureau of Statistics which consist of eight major crimes incidents for the period January 2005 to December 2012 except the data of August 2006 which was not available. Eight major crimes can be compromise in two groups namely crime against person: murder, attempted murder, kidnapping/abduction and crime against property: dacoity, robbery, burglary, cattle theft, other theft.

Principal Component Analysis

According to [15] explanation about the main idea of the PC transformation, PCA is used to retain few (<p) derived variables preserving most of the information provided by the variance of the p random variables. This linear transform has been widely adopted in data analysis and compression [16].

Let X be a vector of p random variables X' = [X_1, X_2, …, X_p **]** having the covariance matrix $\sum$ with eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0.$$

Let the element of X has the following linear combinations

$$Y_j = \alpha'_j X = \alpha_{j1}X_1 + \alpha_{j2}X_2 + \ldots + \alpha_{jp}X_p = \sum_{k=1}^{p} \alpha_{jk}X_k, \quad j = 1,2,\ldots,p$$

With a vector of p components $\alpha_{j1}, \alpha_{j2}, \ldots, \alpha_{jp}$.

Then $Var(Y_j) = \alpha'_j \sum \alpha_j$

$$j = 1,2,\ldots,p \qquad (1.1)$$

$Cov(Y_j, Y_k) = \alpha'_j \sum \alpha_k$

$$j = 1,2,\ldots,p \qquad (1.2)$$

The PCs are those uncorrelated linear combinations $Y_1, Y_2, \ldots, Yp$ whose variances in (1.1) are as large as possible [17]. Emphasis on the variances is given in finding the PCs. First of all we look for a linear combination with maximum variance, such that

$\alpha'_1X = \alpha_{11}X_1 + \alpha_{12}X_2 + \ldots + \alpha_{1p}X_p = \sum_{k=1}^{p} \alpha 1kXk$

Next, we look for a linear combination α'_2X uncorrelated with α'_1X having maximum variance and so on, at the end we reach at kth stage of linear combination α'_kX having maximum variance and also being uncorrelated with α'_1X, $\alpha'_2X, \ldots, \alpha'_{k-1}X$. The kth PC is kth derived variable $\alpha'_k X$. Although upto p PCs could be derived but we restrict our findings till the qth stage ($q \leq p$) when most of the variation in X have been accounted for by q PCs.

Given $Var(Y_j) = \alpha'_j \sum \alpha_{j=\lambda j}$

$$j = 1,2,\ldots,p$$

is the variance of PC which is equal to the corresponding eigen value

The total variance of PCs is considered as the total variance in a data set, which is given below

$$\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \sum_{j=1}^{p} Var(Xj) = \lambda 1 + \lambda 2 + \ldots + \lambda p = \sum_{j=1}^{p} Var(Xj)$$

By standardizing the variables X' = [X_1, X_2,…, X_p **]** of similar scale with mean zero and unit standard deviation, we have the following corresponding standardized variables

$$Z = [Zj = \frac{(Xj - \mu j)}{\sigma jj}]$$

$$j = 1,2,\ldots,p$$

In matrix $Z = (V^{1/2})^{-1} (X - \mu)$

where $V^{1/2}$ is the diagonal standard deviation matrix having te following properties

E(Z)= 0

Cov(Z)=ρ

The eigenvectors of the correlation matrix ρ of X will provide the PCs of Z, having all the properties of X by referring Y_j to the jth PC and (λ_j, α_j) to eigenvalue – eigenvector pair.

Now

The jth PC of the standardized variables Z' = [$z_1, z_2, \ldots, z_p$] can be shown as below

$Y_j = \alpha'_j Z = \alpha'_j (V^{1/2})^{-1} (X - \mu)$,

Such that

$$\sum_{j=1}^{p} Var(Xj) = \sum_{j=1}^{p} Var(Zj) = p$$

$$j = 1,2,\ldots,p$$

Having the following he eigenvalue- eigenvector pairs for ρ

$(\lambda_1, \alpha_1), (\lambda_2, \alpha_2), \ldots, (\lambda_j, \alpha_j)$ with $\lambda_{11} \geq \lambda_2 \geq \ldots \geq \lambda_{p \geq} 0.$

Interpretations of outcomes of Principal Component Analysis

The loading or the eigenvector $\alpha j = \alpha_1, \alpha_2, \ldots, \alpha_p$ shows the importance of the variable for a given PC. The eigenvector with the highest eigenvalue is the most dominant principle component of the dataset (PC_1). It expresses the most significant relationship between the data dimensions [18]. The type of crime components can be found by analyzing the positive and negative coefficients in subsequent components [19]. The information about the weights of original variables when calculating each PC can be found in loading matrix which shows association between PC and original variable [20].

The proportion of variance:

The best explanation of the original variables is obtained by the proportion of variance which is given below

$$\Psi q = \frac{\sum_{j=1}^{p} \lambda j}{P} = \frac{\sum_{j=1}^{p} Var(Zj)}{P}$$

A useful criterion for determining the number of components to be retained in the analysis is called cumulative proportion of explained variance. A good graphical representation of the ability of the PCs to explain the variation in the data is a scree plot [21].

RESULTS AND DISCUSSION

Table 1 explores the different levels of correlations between the crimes. Crime against property and by person are moderately correlated which means variables can be used to predict each other except in case of cattle theft. Moderate correlations are found between crimes against property which means prediction of each crime can be made by other variables. Same is the case with crime against a person where moderate correlation is found between crimes except kidnaping and murder, which are strongly correlated and can be used to predict each other which mean high rate of murder in state is associated to kidnappings. The Gleason-Staelin reduncy measure, phi is 0.63 which indicates moderate inter correlation among variables. But care must be taken in case of Phi value is less than 0.5 [22].

Table 1: Correlation Matrix

Variables	Murder	Attemped_ Murder	Kidnapping_ Abduction	Dacoity	Robbery	Burglary	Cattle_ theft	Other_ theft
Murder	1.0000							
Attemped_ Murder	0.8653	1.0000						
Kidnapping_ Abduction	0.9074	0.6982	1.0000					
Dacoity	0.6541	0.4317	0.7609	1.0000				
Robbery	0.5619	0.4572	0.6611	0.8015	1.0000			
Burglary	0.7223	0.5067	0.8448	0.7426	0.6826	1.0000		
Cattle_theft	-0.1630	0.0113	-0.2210	-0.2910	-0.2883	-0.0692	1.0000	
Other_theft	0.7820	0.6204	0.8937	0.7842	0.7405	0.8386	-0.1770	1.0000

Phi=0.635480 Log(Det|R|)=-9.065131 Bartlett Test=820.39 DF=28 Prob=0.0000

Table 2: Eigenvalues of Correlation Matrix

No.	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	5.352713	66.91	66.91	||||||||||||
2	1.118910	13.99	80.90	|||
3	0.717692	8.97	89.87	||
4	0.362613	4.53	94.40	|
5	0.193277	2.42	96.82	|
6	0.140687	1.76	98.57	|
7	0.080010	1.00	99.57	|
8	0.034099	0.43	100.00	|

Table 2 displays eigenvalues, percent and cumulative percent of explained variance which will help us to decide how many factors (or components) are being retained. As rule of thumb factors having eigenvalues greater than one are sufficient to be retained [12]. However, by considering scree plot in figure 1, it is reasonable to retain first three components as third eigenvalue $\lambda = 0.71$ is approximately close to 1. Thus by retaining first three PCs up to 89.87% of the variability in the total data set can reasonably be explained.

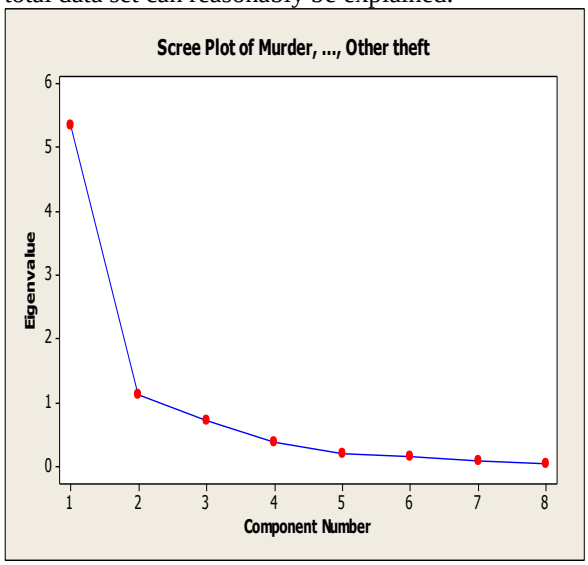

Figure 1: A Scree Plot for Crime Rate

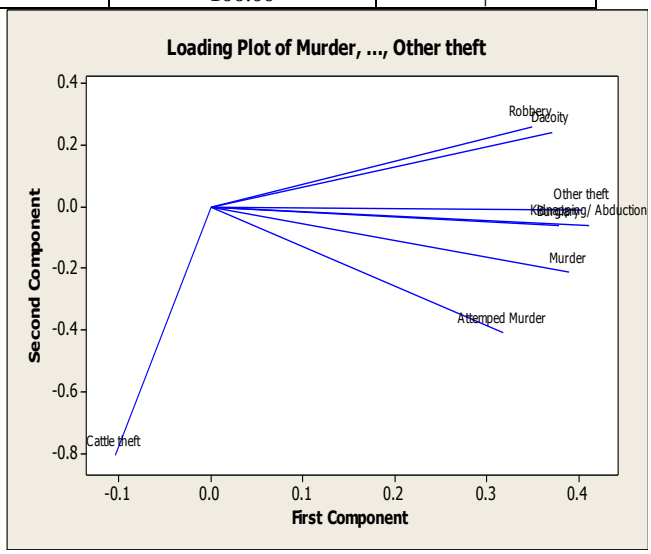

Figure 2: Loading Plot

Figure 2 shows classification of crime according to low and high crime rate by using loading plot where robbery has the highest crime rate and cattle theft with lowest crime rate, moderate positive loading on dacoity, other theft and kidnapping, and small negative loading on murder and attempted murder. From table 3, coefficients in PC (or factor) one shows the relative importance of each crime in forming the factor with negative weights ranging (0.4119 -0.1050) and shows an overall measure of crime in state.

Table 3: Eigenvectors

Variables	Factors							
	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8
Murder	-0.3894	0.2133	0.3704	0.0827	-0.1942	0.1845	-0.3164	-0.6979
Attemped_ Murder	-0.3184	0.4106	0.5108	-0.4131	0.0697	0.0617	0.4261	0.3290
Kidnapping_ Abduction	-0.4119	0.0586	0.0932	0.3340	-0.0435	-0.0917	-0.5886	0.5915
Dacoity	-0.3709	-0.2412	-0.2997	-0.1455	-0.7934	0.0331	0.2379	0.0786
Robbery	-0.3496	-0.2596	-0.3111	-0.6685	0.3858	0.1138	-0.3180	-0.0617
Burglary	-0.3791	0.0616	-0.3338	0.4456	0.3328	0.5511	0.3578	-0.0032
Cattle_theft	0.1050	0.8081	-0.5265	-0.1303	-0.1219	-0.0596	-0.1517	-0.0196
Other_theft	**-0.4035**	**0.0095**	**-0.1340**	**0.1655**	**0.2273**	**-0.7953**	**0.2512**	**-0.2109**

Table 4: Factor Loadings

Variables	Factors							
	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8
Murder	-0.9008	0.2256	0.3138	0.0498	-0.0854	0.0692	-0.0895	-0.1287
Attemped_ Murder	-0.7365	0.4344	0.4327	-0.2487	0.0306	0.0231	0.1205	0.0608
Kidnapping_ Abduction	-0.9530	0.0620	0.0790	0.2011	-0.0191	-0.0344	-0.1665	0.1092
Dacoity	-0.8580	-0.2551	-0.2539	-0.0876	-0.3488	0.0124	0.0672	0.0145
Robbery	-0.8087	-0.2747	-0.2635	-0.4025	0.1696	0.0427	-0.0810	-0.0114
Burglary	-0.8772	0.0652	-0.2828	0.2684	0.1463	0.2067	0.1012	-0.0006
Cattle_theft	0.2430	0.8548	-0.4460	-0.0785	-0.0534	-0.0224	-0.0429	-0.0036
Other_theft	-0.9336	0.0100	-0.1136	0.0996	0.0999	-0.2983	0.0711	-0.0389

Table 5: Bar Chart of Communalities

Variables	Factors								Commun-ality
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	
Murder	|||||||	||	||					||||||||||||||||||
Attemped_ Murder	|||||||||	||||	||||	||				||||||||||||||||||
Kidnapping_ Abduction	||||||||||||||||||								||||||||||||||||||
Dacoity	||||||||||||||	||	||		|||			||||||||||||||||||
Robbery	||||||||||||||	||	||	||||				||||||||||||||||||
Burglary	||||||||||||||		||	||				||||||||||||||||||
Cattle_theft	||	||||||||||||||	||||					||||||||||||||||||
Other_theft	||||||||||||||||||					||		||||||||||||||||||

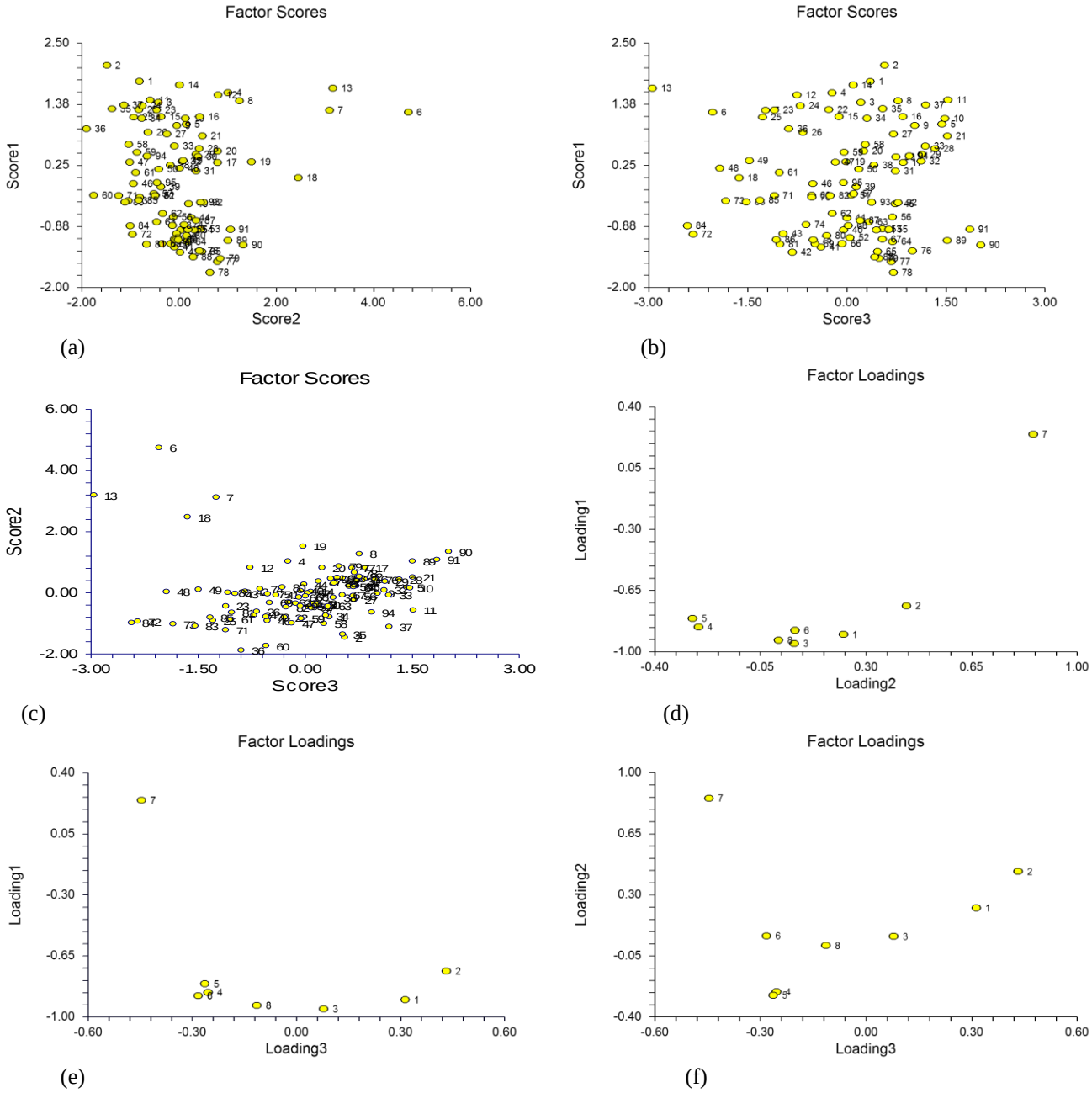

Figure 3: Factor Loading

From table 4 and table 5, it can be seen that factor 1 contains high moderate factor loadings for all crime categories except cattle theft which means the crime rate of these crime categories including crime against property and by person can be represented by this component. Three of these crimes categories (murder, kidnapping/abduction and other theft) have factor loadings of over 0.9 which means only this component reasonably explains over 89.87% of the total variation of each of these crimes. All other remaining factors show very low correlation between crimes in the state.

From figure 3(d-f), we have cattle theft as an outlier. Therefore, corresponding data values in figure 3(a-c) shows no tendency toward cattle theft and has the lowest prevalence of cattle theft crime in the country.

From figure 3d, second PC shows murder, burglary, kidnapping, other theft and dacoity has over all more crime prevalence. From figure 3c, linear dependencies can be observed on third PC, indicating existence of interactions among the retaining PCs. Crimes both against property and against person have equal level of risk in the country.

CONCLUSION

A great reduction in dimensionality is achieved by retaining first three PCs out of eight original crime categories by applying principal component analysis to the crime data obtained from Pakistan Bureau of Statistics. It is worth mentioning that retained PCs explained almost 89.87% of the total variability in the original data set ensuring less loss of the information. Moderate strength of association exists

between crime against property and crime against person except in case of cattle theft. Furthermore, crimes against property are moderately correlated with each other. High correlation exists between kidnaping and murder, and can be used for predicting each other. However, moderate correlation is found between remaining crime categories of crimes against persons.

According to second PC cattle theft crime is detected as an outlier having least crime rate. The second PC classifies the crimes, according to crime rates namely (1) most popular crimes: kidnaping, burglary and other crimes (2) least popular crimes: attempted murder, dacoity and robbery. Solution of complex criminal problems that have bedeviled the country can be addressed by successful implementation of suggested crime patterns. Information provided in this paper can be helpful to local governments by paying more enforcement to the crimes with high crime rate in the state. Furthermore, government's priorities can be setup by following correlation patterns exist between various crimes.

REFERENCES

[1] Fedotov, Y. (2013). Global Study On Homicide 2013: Trends, Contexts, data. United Nations Office on Drugs and Crim (UNODC), Vienna.

[2] Harrendorf, S., Heiskanen, M., and Malby, S. (2010). International Statistics on crime and Justice. European Institute for Crime Prevention and Control, Finland.

[3] Harrington, V. and Mayhew, P. (2002). Mobile Phone Theft. Home Office Research Study No 235. London: Home Office.

[4] Aromaa, K. and Nevala, S. (2004). Crime and Crime Control in an Integrating Europe Plenary presentations held at the Third Annual Conference of the European Society of Criminology, European Institute for Crime Prevention and Control, Finland.

[5] Financial (2011). Nigeria crime. Financial Times, 7/11/2011.

[6] Smit, E. (2013). Spatial Analysis of South African Crime Data. University of the Witwatersrand, Johannesburg, South Africa, Proceedings 59th ISI World Statistics Congress, Hong Kong, pp.5410-5414.

[7] Gillani, S. Y. M., Rehman, H. and Gill, A. R. (2009). Unemployment, Poverty, Inflation and Crime Nexus: Cointegration and Causality Analysis of Pakistan. Pakistan Economic and Social Review, 47(1),79-98.

[8] Pakistan Crime Monitor, (April 2013). Free and Fair Election Network (FAFEN), Isla mabad-Pakistan.

[9] Ahaman, B. (1967). An Analysis of Crimes by the Method of Principal Components. Journal of the Royal Statistical Society, Series C (Applied Statistics), 16(1),17-35.

[10] Bello, Y., Batsari, Y. U., and Charanchi, A. S. (2014). Principal Component Analysis of Crime Victimizations in Katsina Senatorial Zone. International journal of scienc and techniology, 3(4),192-202.

[11] Salvati, L., Di Bartolomei, R., Rontos, K. and Bisi, S. (2012). Crime and the (Mediterranean) city: Exploring the geography of (in) security in Rome, Italy. International Journal of Latest Trends in Finance and Economic Sciences, 2(1),56-73.

[12] Malhotra N. and Dash S. (2011), Marketing Research-An applied orientation, Sixth Edition, Pearson Publication.

[13] Sandbhor, S., Choudhary, S., Arora, A. and Katoch, P. (2014). Identification of Factors Leading to Construction Project Success Using Principal Component Analysis. International Journal of Applied Engineering Research, 9(17),4169-4180.

[14] Gulumbe, S. U., Dikko, H. G. and Bello, Y. (2013). Analysis of Crime Data using Principal Component Analysis: A case study of Katsina State. CBN Journal of Applied Statistics, 3(2),39-49.

[15] Jolliffe, I. T. (2002). Principal Component Analysis 2nd edn. New York: Springer-verlag.

[16] Banerjee, A. (2012). Impact of Principal Component Analysis in the Application of Image Processing. International Journal of Advanced Research in Computer Science and Software Engineering, 2(1), http://www.ijarcsse.com/docs/papers/january2012/V2I1052.pdf visited on 02/09/2014.

[17] Richard, A. J. And Dean, W. W. (2001). Applied Multivariate Statistical Analysis, 3rd edn. New Delhi: Prentice Hall.

[18] Jeong, D. H., Ziemkiewicz, C., Ribarsky, W. and Chang, R. (2008). Understanding Principal Component Analysis Using a Visual Analytics Tool. Charlotte Visualization Center, UNC Charlotte.

[19] Rencher, A. C. (2002). Methods Multivariate Analysis, 2nd edn. New York: Wiley.

[20] Fang, Y. (2011). Multivariate Methods Analysis of Crime Data in Los Angeles Communities. University of California, Los Angeles.

[21] Cattell, R. B. (1966). The Scree Test for the number of Factor. Multivariate Behavioral Research, 1,245-276.

[22] Usman, U., Yakubu, M. and Bello, A. Z. (2012). An Investigation on the Rate of Crime in Sokoto State Using Principal Component Analysis. Nigerian Journal of Basic and Applied Science, 20(2),152-160.