American Journal of Theoretical and Applied Statistics
2015; 4(4): 258-263
Published online June 3, 2015 (http://www.sciencepublishinggroup.com/j/ajtas)
doi: 10.11648/j.ajtas.20150404.15
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Principal Component and Principal Axis Factoring of Factors Associated with High Population in Urban Areas: A Case Study of Juja and Thika, Kenya

Josephine Njeri Ngure[1], J. M. Kihoro[1], Anthony Waititu[2]

[1]School of Mathematics, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
[2]Co-ooerative University College of Kenya, Nairobi, Kenya

Email address:
ngurejosephine@gmail.com (J. N. Ngure)

To cite this article:
Josephine Njeri Ngure, J. M. Kihoro, Anthony Waititu. Principal Component and Principal Axis Factoring of Factors Associated with High Population in Urban Areas: A Case Study of Juja and Thika, Kenya. *American Journal of Theoretical and Applied Statistics.* Vol. 4, No. 4, 2015, pp. 258-263. doi: 10.11648/j.ajtas.20150404.15

Abstract: Development in the world / a country today is being influenced by the population in urban areas as a result of which living standards rise in all parts of the country despite the rural areas. The main goal of our government today is to balance development of urban and rural areas of Kenya so that no areas are left behind as others head forward in terms of development. In this research, PCA and PAF methods of factor reduction were applied. PCA is a widely used method for factor extraction. Factor weights are computed in order to extract the maximum possible variance, with successive factoring continuing until there is no further meaningful variance left. The factor model is then rotated for analysis. PAF restricts the variance that is common among variables. It does not redistribute the variance that is unique to any one variable. Parallel analysis, catell's scree test criterion and Eigen value rule were applied. Results indicated that parallel analysis was generally the best the scree test was generally accurate while the Kaiser's method tended to overestimate the number of components. In this research, business and employment were deduced as major factors associated with high population in the two towns. Amenities like telephone networks, markets were also associated with high population in the two towns. I recommend the Kenyan government to apply the knowledge of PCA and PAF to determine the major reasons associated with high population in other major urban areas (towns and cities) especially according to 2009 population and housing census results so as to assist in allocation of revenue in the now current devolution system of government. This will ensure no areas (counties) are left behind in terms of development. The government should strive to provide social amenities and utilities in the rural areas. It should also provide jobs to the citizens in the rural areas so as to prevent very high increase in urban areas. The people in rural areas can also hold vocational training on self employment being headed by the government. PAF method demonstrated better results than the PCA since it took good care of measurement errors. PAF method was also able to recover weaker factors than PCA could. PAF removed the unique and error variance and so its results were much more reliable. PAF was also preferred because it accounted for the co-variation whereas PCA accounted for the total variance.

Keywords: Principal Component Analysis, Principal Factor Analysis or Common Factor Analysis or Principal Axis Factoring, Factor Analysis, Kaiser Meyer Olkin

1. Introduction

Background Information

An urban area is normally built up with roads, houses, shops, offices, entertainment centers and public buildings. In most cases, an urban area is a relatively large to huge city, often with buildings some over 10 stories high, no land for farming and it has a much denser population. Rural areas on the other hand are areas with low population per square kilometer, with a lot of land for farming and open space. People move from rural to urban areas mainly to look for employment opportunities, better education facilities, better business opportunities and better security. According to Population Reference Bureau, World population highlights 2007 indicated that the world was / is on the verge of a shift from predominantly rural to mainly urban. It had predicted that in 2008, more than half the world's people will live in urban areas. By 2030, urban dwellers will make up roughly 60% of the world's population. It also indicated that globally, all future population growth will take place in cities, especially in Asia, Africa and Latin America. By 2030, Africa and Asia will account for almost seven out of 10 urban dwellers globally and poor people will make up a large part of future urban growth. By 2011 (The fourth Kenya Economic Update 2011 titled "Turning the Tide in Turbulent Times"), at least 30% of Kenyans lived in cities (with at least 8.3% annual growth rate) and most Kenya's population growth was urban. While total population was expected to double by 2045, the urban population is expected to be more than quadruple. By 2033, half of Kenya will be residing in urban areas. With more people in the same space, the cities will be more and bigger and Kenyan cities of 100,000 people and above will grow from around 21 to 37 in 2020.

Thika and Juja are both found in Kiambu County. Juja Constituency falls under Thika District in Central Province of Kenya. According to the 1999 census, Thika District (Thika East and Thika West districts, urban) had a population density of 329 people while Juja constituency had a population density of 345.3. According to Kenya open data survey 2014, Thika District (urban) had a population density of 1467.4 while Juja constituency had 652.04.According to 2009 population and housing census, Juja was ranked number three out of 10 most populated constituencies in the country.

2. Review of Previous Studies

Zainab Gimba, PhD (2012) focused on the cause and effects of rural-urban migration where the writer did a survey in the town amongst 150 respondents and found out that the major cause of high population in urban areas was as a result of migration with the causes being to search for education, employment and business opportunities, while in rural areas, poverty, unemployment, famine and inadequate social amenities led to low population.

Milner and Wimberley (October 1980) administered the Child Abuse Potential Inventory to 65 abusing and 65 matched non-abusing parents. An item analysis indicated that 77 of the 160 Inventory items significantly discriminated between abusers and non-abusers. A discriminant analysis which employed the 77 significant items, correctly classified 125, or 96%, as abusers or non-abuser. Identical results were obtained from a stepwise multiple regression analysis. PAF with oblique promax rotations yielded a seven-factor solution with relatively specific dimensions. The seven factors were: distress, rigidity, child with problem, problem from family and others, unhappiness, loneliness, and negative concept of child and of self. Further analysis revealed that the rigidity, unhappiness, and distress factors were the most meaningful dimensions in the understanding of why some people abuse children.

PCA has been applied in Anthropological studies (Chen et al., 2011). The research involved study of gene expression and it concluded that PCA has a population genetics interpretation and can be used to identify differences in ancestry among populations and samples, though there are limitations due to the dynamics of micro evaluation and historical processes.

A self-report inventory for the assessment of mindfulness skills was developed by Baer et al. (2004) and its psychometric characteristics and relationships with other constructs were examined. Three samples of undergraduate students and a sample of outpatients with borderline personality disorder were the participants. Based on discussions of mindfulness in the current literature, four mindfulness skills were specified: observing, describing, acting with awareness, and accepting without judgment. Scales designed to measure each skill were developed and evaluated. Results after PFA showed good internal consistency, test-retest reliability and a clear factor structure. Most expected relationships with other constructs were significant. Findings suggested that mindfulness skills are differentially related to aspects of personality and mental health, including neuroticism, psychological symptoms, emotional intelligence, alexithymia, experiential avoidance, dissociation, and absorption.

3. Methodology

3.1. Data Collection

The research was on an unbiased and valid sample collected from Juja and Thika towns in Kenya which was to be a representative of the whole population. The target group for collection of data was a family head or representative who could avail the required information as was needed. Questionnaires were administered to appropriate respondents who could provide necessary response to the set questions. There was also use of interviews in case of aged and semi-illiterate respondents and call backs in case of missing information. No secondary data was used in this research.

3.2. Review of Principal Component Analysis

A set of n-dimensional vector samples X= $\{x_1, \ldots x_m\}$ is transformed into another set Y = $\{y_1, \ldots y_m\}$

Of the same dimensionality, but y-s have the properties that most of their information content is stored in the first few dimensions. So, it's possible to reduce the data set to a smaller number of dimensions with very low information loss Smith (February 26 2002).

The transformation is based on the assumption that high information corresponds to high variance. **X** is transformed into **Y** as a matrix computation **Y**= **A. X** choosing **A** such that **Y** has the largest variance possible for a given data set. The single dimension **Y** obtained in this transformation is called the first principal component. This component is an axis in the direction of maximum variance.

3.3. Review Principal Axis Factoring

PAF (most widely used method in factor analysis) a type of exploratory factor analysis, restricts the variance that is common among variables. i.e. it does not redistribute the variance that is unique to any one variable. A common factor is an abstraction / hypothetical dimension that affect at least two of the variables. It is assumed that there is one unique factor for each variable, a factor affecting that variable but does not affect any other variables. It is also assumed that the q unique factors are uncorrelated with one another and with the common factors. It is the variance due to these unique factors that is excluded from the PFA.

Supposing the observed variables are $X_1, X_2, \ldots X_n$ the common factors are $F_1, F_2, \ldots F_m$ and the unique factors are the variables can then be expressed as linear functions of the factors $U_1, U_2, \ldots U_n$ the variables can then be expressed as linear functions of the factors: equation 20 below

$$X_1 = \alpha_{11}F_1 + \alpha_{12}F_2 + \alpha_{13}F_3 + \ldots + \alpha_{1m}F_m + \ldots \alpha_1 U_1$$

$$X_2 = \alpha_{21}F_1 + \alpha_{22}F_2 + \alpha_{13}F_3 + \ldots + \alpha_{2m}F_m + \ldots \alpha_2 U_2$$

$$X_3 = \alpha_{31}F_1 + \alpha_{32}F_2 + \alpha_{33}F_3 + \ldots + \alpha_{3m}F_m + \ldots \alpha_3 U_3$$

Each of these equations is a regression equation; factor analysis seeks to find the coefficients which best reproduce the observed variables from the factors. The coefficients α_{11}, α_{12}, α_{13}, …,α_{nm} are the loadings in the same way as regression coefficients. In the model 20 above, α_{11} is the loading for variable X_1 on F_1, α_{11} is the loading for variable X_2 on F_2, and so on. When the coefficients are correlations, i.e., when the factors are uncorrelated, the sum of the of the squared loadings for variable X_1 namely

$[\alpha^2_{11} + \alpha^2_{21} + \alpha^2_{31} + \ldots \alpha^2_{n1}]$ shows the proportion of the variance of all the variables which is accounted for by that factor.

4. Data Analysis and Results

4.1. Principal Component Analysis

The first thing was to check whether the sample was big enough for factor analysis. This was done by checking the KMO measure of sampling adequacy (KMO test). Sampling adequacy predicted if the data was likely to factor well, based on correlation and partial correlation. The diagonal elements on the Anti-image correlation matrix were the KMO individual statistics for each variable. In this case, the total value of KMO was found to be 0.97 which was actually a superb value and so factor analysis was appropriate for this data (Sofroniou N. and Hutcheson, 1999). The KMO measure of sampling obtained was 0.907 and Bartlett's test of sphericity was significant at 780 degrees of freedom. The determinant of the correlation matrix was 0.0000526 which was greater than 0.00001 and showed no multicollinearity in the variables.

***Table 1.** KMO and Bartlett Tests Results*

KMO and Bartlett test	
Statistic	**Value**
KMO	0.97
Bartlett Test of Sphericity	Approximately .Chi-Square 1106.65 Df 780 sig 0.000

The KMO measure of sampling obtained was 0.907 and Bartlett's test of sphericity was significant at 153 degrees of freedom. The determinant of the correlation matrix was 0.0000526 which was greater than 0.00001 and showed no multicollinearity in the variables. All these measure indicated that the data was fit for factor analysis. Initially the total variance is 40, which is equal to the number of standardized variables (40 variables). This is because for standardized data, the variance of each standardized variable is 1. The cumulative proportion of variance explained criterion suggests we retain as many PCs as are needed in order to explain approximately 80-90% of the total variance, and here the number of PCs needed was 17.

Kaiser's rule suggests to retain as many PCs as are those whose variance is larger than the average variance. According to this criterion,retain the number of PCs that have a variance (eigen values) larger than 1 since we had standardized the variables. In this case the PCs were observed to be 10. The variance was above 1 for principal components 1, 2, up to 10 (which have variances 10.61, 3.12 up to 1.06). Therefore, using Kaiser's criterion, we would retain the first 10 principal components. The most change in slope in the scree plot Figure 1 below occurs at component 5, which is the "elbow" of the scree plot. Therefore, it could be argued based on the basis of the scree plot that the first 5 components should be retained. This figure showed the variances represented by the first 10 principal components as explained by the eigen values.

The Scree plot for the three extraction methods Figure 2 below (Kaiser's rule/the Eigen value criterion, scree plot criterion and the parallel analysis) was as shown The acceleration factor indicated where the elbow of the scree plot appears. It corresponds to the acceleration of the curve i.e. the second derivative. The optimal coordinates are the extrapolated coordinates of the previous Eigen value that allow the observed eigen value to go beyond this extrapolation (Stephens, 20th April 2009). Here, Eigen value criterion retained 10 factors as seen earlier, parallel analysis retained 7 factors while the scree plot criterion retained 7 factors.

After varimax rotation, the first rotated component had high loadings on places to buy clothes (0.83), telephone networks(0.80), supermarkets(0.77), market(0.75) and so was named factor "social amenities", the second had high loadings on land for farming (0.75), less theft (0.67) and being close to family members(0.54) and so led to the factor "social life". The third rotated component had high loadings on one self and the spouse being employed in Juja (0.59 and 0.62 respectively), one being a business person in Juja (0.6) and so was named "economic activities ", the fourth had high loadings on cheap rental house (0.51), one being employed in Thika (0.65), a spouse being employed in Thika (0.57), . So, the third and fourth could be combined to the factor "economic activity". The fifth was had high loadings on one being a business person in Thika (0.77), one being a business person in Nairobi (0.73), and this being an ancestral home (0.62) and so was also named "economic activity" . The first five rotated components explained a total of 50% of the variance.

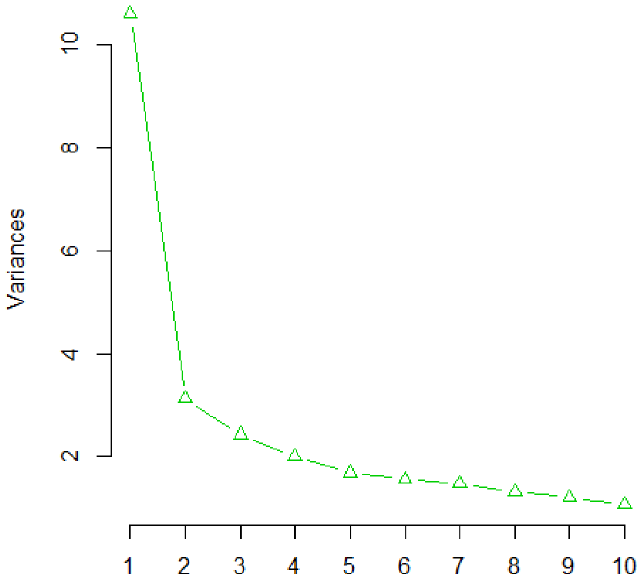

Figure 1. *Scree plot ofvariances against components.*

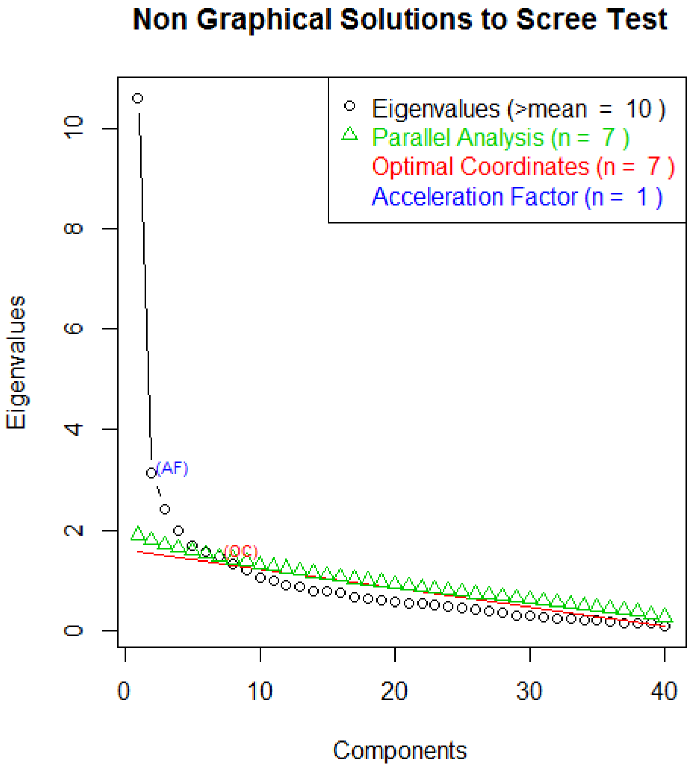

Figure 2. *Scree plot for the three extraction methods.*

4.2. Principal Axis Factoring

PAF sought the least number of factors which caould account for the common variance (correlation) of a set of variables. Here, the variance due to unique factors is normally eliminated by replacing the 1's on the main diagonal of the correlation matrix with estimates of the variable's communalities (amount of the variable's variance that is accounted for by the components or factors). The determinant of the correlation matrix was 0.0000957 which was greater than 0.00001 (of the recommended value) which indicated that factor analysis was possible for this data. The KMO value was found to to be 0.80907 which was a superb value and which also indicated that the data was fit for factor analysis.. The Bartlett test of Sphericity tested the null hypothesis that the correlation matrix was an identity matrix. The test was significant at p<0.001.

Table 2. *Kmo and Bartlett Test Results*

KMO and Bartlett test	
Statistic	**Value**
KMO	0.80907
Bartlett Test of Sphericity	Approximately Chi-Square 3957.4 Df 780 sig 0.000

In the next steps, the data was first converted to numeric (by forming a matrix containing all items). The data set was arranged such that the observations were the rows and the variables were in the columns. The Eigen value criterion was used and eigen values greater than 1 were extracted. After 23 iterations, there were 7 Eigen values greater than 1 which reflected the proportion of the variance explained by the factors. This showed that we had 7 common factors .The standardized loadings based upon the correlation matrix were also displayed. The first factor had variables, by looking at the standardized factor loadings based on the correlation matrix, plenty of supermarkets (0.79),good places to buy clothes for my family (0.81), plenty of clubs and hotels (0.73), primary schools (0.77) and nearness to different telephone networks (0.79) loading highly on this first factor and so this factor was named “social amenities”. The second factor had high loadings on land (0.67), being close to friends (0.58), nearness to police station (0.49), less theft (0.45), general hospitals (0.42), dispensaries (0.41), private hospitals (0.36) and so was named “social utilities” since these are necessities that make life comfortable. The third factor had high loadings on employment in Juja both for one self and the spouse (0.56 and 0.51), spouse being employed in Nairobi (0.42), and business person in Juja (0.52) and so was named “economic activities”. Factor 4 had high loadings on nearness to capital center (0.51), employment in Thika both for one self and the spouse (0.55 and 0.46 respectively), cheap rental houses (0.45) and nearness to Thika road (0.41) and so was named “infrastructure”. The fifth factor had high loadings on business person in Thika (0.77), business person in Nairobi (0.73) this being an ancestral home (0.62) and so was named "business opportunity" which is actually a sub-branch of “economic activity” factor. The rest of the factors were a repetition of the previous.

The proportion explained by the first factor was 40% the second 14% and so on. The cumulative proportion explained by the first five factors was 84%.

Table 3. *Comparison of results.*

Principal Component Analysis	**Principal Axis Factoring**
The method does not take into account the measurement errors. It assumes that each variable is perfectly reliable and that's why the method produced initial communalities as 1's.	Takes into account the measurement errors (the variance not attributable to the factor an observed variable represents) and thus does not produce initial communalities as one.
The method is not able to discover weaker factors.	The method is able to recover weaker factors from weaker the factor loadings. (Dodou and Winter, April 2012)
Its goal was to determine the structure.	Its goal was to discover the structure and not to determine it.
The method extracts the unique variance only. It did not take into account the error variance.	PFA removed the unique and error variance and so its results were much more reliable.

5. Discussions, Conclusions and Recommendations

The rate of rural-urban migration is alarming in the recent years and its effects are not only felt in the urban areas alone but also felt by the source regions since the facilities in the destination regions are overstretched and the source regions are virtually deserted. Unless the government provides the basic necessities of life to the rural areas and provide the productive youths in the rural areas with employment, highlighted as one of the factors associated with high population in the two towns, people will continuously drift to the urban areas from the rural areas in search of better life and employment. CFA outperformed PCA in determining the major factors associated with the high population in the two urban areas since it took into account the measurement errors, removed the unique and error variance which made its results more reliable. PCA based on the covariance matrix and PCA based on the correlation matrix yielded similar results as read in Raschka (13th April 2014) since a correlation matrix is simply a standardized covariance matrix.

I recommend more research and application to be done in other areas of factor reduction (factor analysis) besides the two methods used in this research. Other methods like alpha factoring, maximum likelihood, image factoring could also be applied when one is conducting factor reduction. I also recommend the Kenyan government to apply the knowledge of PCA and CFA to determine the major reasons associated with high population in major urban areas (towns and cities) especially according to 2009 population and housing census results so as to assist in allocation of revenue in the now current devolution system of government. This will ensure no areas (counties) are left behind in terms of development. In view of the factors associated with high population in urban areas, the government should strive to provide social amenities and utilities in the rural areas. It should also provide jobs to the citizens in the rural areas so as to prevent very high increase in urban areas. The people in rural areas can also hold vocational training on self employment being headed by the government.

References

[1] Baer, Ruth A., Gregory T. Smith University of Kentucky, Gregory T. Smith University of Kentucky and K. Kristin B. Allen Comprehensive Care Center, Lexington (2004), 'Assessment of mindfulness by self-report .the Kentucky inventory of mindfulness skills', Sage publication .

[2] Bartlett, M. S. (1954), A note on the multiplying factors for various chi square approximations, New York: HarperCollins.

[3] Catell, R. B. (1966), The scree test for number of factors, Multivariate Behavioral Research, published online.

[4] Chajewski, Michael (2008), 'Item analysis with alpha standard error and principal axis factoring for continuous variable scales (with plots).', The college Board Research and Development, Journal of Psychometrics.

[5] Chen, Xingdong, Chao Chen and Li Jin (2011), 'ministry of education key laboratory of contemporary anthropology, fudan university, shanghai china', Journal of Scientific Research.

[6] Choi, N, D Fuqua and W Griffin, B (2001), Exploratory analysis of the structure of scores from the multidimensional scales of perceived self efficacy, Sage publications.

[7] Davis, James E, Allan Shepard, Nancy Stanford and L.B Rogers (1974), 'Application of pca to combined gas chromatographic-mass spectromic data: department of chemistry, Purdue university, west Lafayette, ind.47907, anal chem.'.

[8] Dodou and Winter (April 2012), 'Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. department of biomechanical engineering, faculty of mechanical, maritime and materials engineering, delft university of technology, mekelweg 2, 2628 cd delft, the Netherlands.', Journal of Applied Statistics Vol. 39 .

[9] George J. Knafl, PhD Professor & Senior Scientist knaflg@ohsu.edu (2000), 'Current topics in statistics for applied researchers factor analysis, Oregon health and science university'.

[10] J.L, Horn (01/06/1965), 'A rationale and test for the number of factors in factor analysis, volume 30, issue no.2', Journal of Psychometrika.

[11] Joost C. F. de Winter, Dimitra Dodou (2012), Common factor analysis versus principal component analysis: a comparison of loadings by means of simulations, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology Mekelweg 2, 2628 CD Delft, The Netherlands. E-mail: j.c.f.dewinter@tudelft.nl.

[12] Kim, J.O and C.W Mueller (1978a), Introduction to factor analysis: what it is and how to do it, Newbury park, CA: sage publication.

[13] Milner, J. S. and Wimberley (October 1980), 'Prediction and explanation of child abuse', Journal of Clinical Psychology.

[14] Novembre, John and Matthew Stephens (18th September 2009), 'Interpreting principal component analysis of spatial population genetic variation', Nature publishing group.

[15] R, Hubbard and Allen S.J (1987), 'An empirical comparison of alternative methods for principal components extraction', Journal of Business Research, 15, 173-190.

[16] Raschka, Sebastian (13th April 2014), 'Implementing a principal component analysis (pca) in python step by step'.

[17] Shlens, Jon (25th March 2003), A tutorial on principal component analysis, derivation, discussion and singular value decomposition.

[18] Smith, Lindsay I (February 26 2002), A Tutorial on Principal Component Analysis. Sofroniou N. and Hutcheson, G.D (1999), The Multivariate Social Scientist: an introduction to generalized linear mmodel., Sage publications.

[19] Stevens, J (1996), Applied multivariate statistics for the social (3rd edn). Tabachnick, B.G and L.S. Fidell (2001), Using multivariate statistics (4th edn)., New York: HarperCollins.

[20] Taylor and Francis Online (1974), 'Application of pca to multitemporal landsat data', International journal of remote sensing.

[21] W.R, Zwick and Velicer W.F (1986), 'Comparison of five rules for determining the number of components to retain', Psychological Bulletin, 432-442.

[22] Zainab Gimba, PhD, M..K. (2012), 'Department of banking and finance, ramat polytechnique. cause and effects of rural-urban migration in borno state', Asian journal of Business and Management Sciences.