African Journal of Mathematics and Computer Science Research Vol. 2(5), pp. 088-091, June, 2009
Available online at http://www.academicjournals.org/AJMCSR
© 2009 Academic Journals

Full Length Research Paper

Statistical analysis of dependence structure of improved cassava varieties in Nigeria

Nwabueze Joy Chioma

Department of Statistics, Abia State University, Uturu, Abia State, Nigeria.

Accepted 20 May, 2009

Multivariate methods were used to analyze a set of data on the proximate compositions of *fufu* flours processed from 43 different cassava mosaic disease (CMD) resistant varieties from National Root Crop Research Institute (NRCRT) Umudike Nigeria. The factor analysis reveals that three factors accounted for 77.8% of the total variables in the data. Factor 1 has eigenvalue of 3.082, factor 2 has an eigen 1.257 and factor 3 has an eigenvalue of 1.005. The scree plot of the data also showed that 3 factors were retained in the study.

Key words: Multivariate statistics, factor analysis, eigenvalue, scree plot, cassava varieties.

INTRODUCTION

Cassava (*Manihot esculenta* Crantz) one of the main agricultural product is a perennial woody shrub with edible root. It is an important staple food crop for millions of people in the tropical areas of African, Asia and Latin America (Rao and Hahn, 1984). It is a staple food whose roots are processed into *garri, fufu*, chips and other fermented products while the leaves equally serve as very nutritional foliage for domestic animals. It also serves as a very vital industrial raw material in the form of starch, chips, pallets, unfermented flours, etc.

The collaborative study on cassava in Africa (COCA) revealed that between 1961 and 1999, total cassava production in Africa nearly tripled from 33 million tones per year from 1995 to 1999 in contrast to the more moderate increase in Asia and Latin America (Nweke et al., 2002). Nigeria is currently the largest producer of cassava in the world with an annual output of over 45 million tones of tubers roots (Anga, 2008). As a result, many improved cassava varieties have been discovered. The cassava varieties were bred for high yield (high dry matter content), pest disease resistance, good product quality and early maturity.

Improved cassava varieties for pest disease resistance are those improved cassava varieties capable of resisting the attack of common cassava disease known as cassava mosaic disease (CMD), a viral disease transmitted by a white fly vector (IITA, 2005). This CMD has been a leading constraint in cassava production in Africa. Where yield losses can be as high as 100%. Researches have been going on for alternative uses of these improved cassava varieties.

Today, cassava has become an important bio-fuel crop in addition to its traditional role, its values as a fuel commodity.

There are various on going cassava fuel ethanol refinery projects in Nigeria that is either under construction or have reached final investment stage. The first cassava to glucose factory in Africa is already up and running in Ogun State, Nigeria. It has an installed capacity of four tones per hour and 100 tones per day (Anga, 2008). These projects are all attributed to the facilitation role of the penitential cassava initiative.

In this study, factor analysis was carried out on a set of data on proximate compositions of *fufu* flour produced from 43 CMD resistant varieties planted in Umudike, Nigeria (Etudaiye et al., 2009). The proximate composition of *fufu* flour moisture, which is used for this study include carbohydrate and dry matter. Given the cost and time involved in measuring all the nutritional compositions of cassava varieties, in this part of the world, this study carrying out this task by subjecting the data to factor analysis. It is important to know which of these proximate compositions of the flour are greatly accounted for *y* the varieties and possibility given the percentage contribution of such variables.

Factor analysis is used as a tool for reducing a large set of variables to a more meaningful smaller set of variables (Crawford and Lomas, 1980). It is a statistical

*Corresponding author. E-mail: teeubueze@yahoo.co.uk. Tel: +2348058810792.

tool that can be used to analyses interrelationships among a large number of variables and to explain these variables in terms of their common understanding dimensions (factors) with a minimum less of information. Factor analysis has had a dual development beginning indirectly with the pioneer of Pearson (1927).

Odimegwu (1999) carried out a research work on family planning attitudes and use in Nigeria using factor analysis. The result showed that the respondents who associated standard of living (factor 1) were more likely to agree to practicing contraception. Other researchers such as Roger et al. (2006) used factor analysis on papers in journal of advanced nursing. Their results revealed that one hundred and twenty-four papers were retrieved. This work investigated the contribution of each composition to the variability of the entire data so as to know which composition contributed maximally. This study also revealed the structure of the data.

Experimental design

Secondary data collected from research work done on *fufu* flour processed from 43 different cassava mosaic diseases (CMD) resistant varieties planted at National Root Crop Research Institute (NRCRI) Umudike, Nigeria (Etudaiye et al., 2008) was used for this work. The amount in percentage of the proximate composition of the *fufu* flours were measured, recorded and used for this work.

Theoretical framework

The model for factor analysis as given by Onyeagu (2003):

$$x_{(px1)} - \mu_{(px1)} = L_{(pxm)}\, F_{(mx1)} + \sum{}_{(px1)} \qquad (1)$$

Where x is the observable random vector with ρ of the ith variable on the ith factor *F* are *m* unobservable random variables called the common factors and Σ are ρ additional sources of variation called error or sometimes specific factors.

The portion of the variance of the ith variable contributed by the *m* common factors is called the ith communality. The portion of the variance of the ith variables $\operatorname{var}(x_i) = \sigma_{ii}$ due to the specific factor is called uniqueness or specific variance. Denoting the ith communality by h_i^2, we have that

$$\sigma_{ii} = h_i^{\,2} + \psi_i \qquad (2)$$

Where σ_{ii} is the variance of the ith variance x$_i$, $h_i^{\,2}$ is the communality of the ith variance and ψ_i is the specific variance.

$$h_i^{\,2} = L^2{}_{i1} + L^2{}_{i2} + \cdots + L^2{}_{im} \qquad (3)$$

Where L_{ij} is the loading of the ith variance on the jth factors? With the assumptions that

$$E(F) = 0_{mx1} \text{ Or}$$

$$\operatorname{cov}(F) = E(FF') = 1_{mxm}$$

$$E(\Sigma) = 0_{px1}$$

$$\operatorname{cov}(\Sigma) = E(\Sigma\Sigma') = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & \psi_p \end{bmatrix} \qquad (4)$$

and that F and Σ are independent so that

$$\operatorname{cov}(\Sigma, F) = E(\Sigma\, F') = 0_{pxm}$$

Factor analysis using principal component method

The covariance matrix can be factored out using the spectral decomposition theorem which states that

$$\Sigma = \lambda_1\, e_1\, e_1' + \lambda_2\, e_2\, e_2' + \cdots + \lambda_p\, e_p\, e_p'$$

$$= \left[\sqrt{\lambda_1\, e_1}\sqrt{\lambda_2\, e_2'} \cdots \sqrt{\lambda_p\, e_p'}\right] \begin{bmatrix} \sqrt{\lambda_1\, e_1'} \\ \sqrt{\lambda_2\, e_2'} \\ \vdots \\ \sqrt{\lambda_p\, e_p'} \end{bmatrix} \qquad (5)$$

Where $(\lambda_i\, e_i)$ is the eigenvalue – eigenvectors pair of Σ and $\lambda_i \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and

$$\Sigma_{pxp} = L_{pxp}\, L_{pxp} + 0_{pxp} = LL^T \qquad (6)$$

The specific factors are now included in the model and their variances are taken to be the diagonal elements of $\Sigma = LL^T$. The approximation becomes.

$$\Sigma = LL^1 + \psi$$

Where $\psi = \sigma_{ii} = \sum_{j=1}^{m} L^2{}_{ij} \quad i = 1, 2, \cdots, p$

Table 1. The correlation matrix of the variables.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1							
X_2	0.162						
X_3	0.238	0.071					
X_4	0.348	0.365	0.498				
X_5	0.000	0.169	0.227	0.433			
X_6	0.000	0.376	0.158	0.308	0.000		
X_7	0.000	0.162	0.238	0.349	0.000	0.000	

Table 2. Total variance explained.

Component	Eigenvalues	% of Variance	Cumulative %
1	3.082	45.459	45.459
2	1.257	17.951	63.410
3	1.005	14.361	77.771
4	0.753	10.758	88.528
5	0.442	6.312	94.841
6	0.361	5.159	100.000
7	2.777×10^{-07}	3.968×10^{-06}	100.000

The communalities are estimated as

$$h^2{}_j = L^2{}_{i1} + L^2{}_{i2} + \cdots + L^2{}_{im} \qquad (7)$$

Where the factor loading L_j' are estimated as

$$L_{ij} = \sqrt{\hat{\lambda}_i \, \hat{e}_{ij}} \qquad (8)$$

Where the contribution to the sample variance s_{ii}. From the first common factor is $L^2{}_{i1}$. The contribution to the total sample variance $s_{11} + s_{12} + \cdots + s_{pp} = tr(s)$ from the first common factor is then

$$L^2{}_{11} + L^2{}_{12} + \cdots + L^2{}_{p1} = \left(\sqrt{\hat{\lambda}_1}\, \hat{e}_1\right)' \left(\sqrt{\hat{\lambda}_1}\, \hat{e}_1\right) = \hat{\lambda}_1 \qquad (9)$$

Since the eigenvalue of total sample variance due to j[th] factor is $\frac{\lambda_j}{p}$.

RESULTS AND DISCUSSION

Since factor analysis depends on correlation analysis, we start by giving the correlation matrix of the variables x_1 to x_7. The results are summarized in Table 1. The correlations matrix was used to asses the sampling adequacy using Kaiser-Mayer-Olkin measure of sampling adequacy (KMO - Test). A KMO value of at least 0.6 is recommended to proceed with factor analysis. We obtained a KMO value of 0.703; we therefore proceeded with factor analysis.

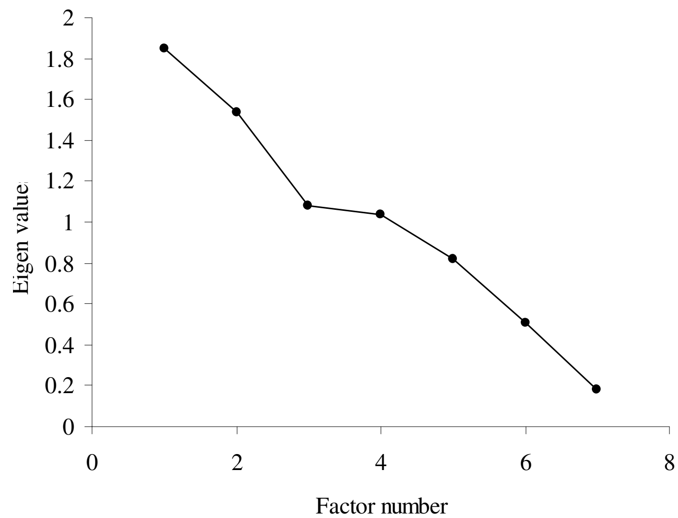

Figure 1. Scree plot.

Number of factors to retain

The result of the factor analysis is summarized in Table 2. The number of factors to be retained in a factor analysis is guided by Kaiser's rule which suggests that those factors with eigenvalues greater than one are retained as the major factors in factor analysis. Based on the eigenvalue of 3.082 for factor 1, factor 2 has eigenvalue of 1.257 and factor 3 has eigenvalue of 1.005. The other factors have eigenvalues less than one. We therefore retained the first three factors in our work. These three factors retained account for about 77.77% of the total variance explained. A scree plot of the variables was also made which is a plot of the eigenvalues along the y – axis against the number of principal components along the x – axis. To know the number of factors to retain, we follow the graph until at a certain point where it will level off; we now retain the number of factors up to that point. From the scree plot on Figure 1, it is clear that 3 factors should be retained.

Factor loading of commonalities

The rotated factor loading and communalities of our data was obtained using varimax. Table 3 summarizes the result. From the Table, factor 1 is made up of the variables x_1, x_5, x_6 and x_7 because of their high loadings. For factor 2, variables x_2 and x_3 have high loading, factor 3 is explained only by variable x_4. We adapted the recom-

Table 3. Retated factor loading and communalities.

Variance	Factor 1	Factor 2	Factor 3	Communalities
X_1	-0.956	-0.044	0.047	0.918
X_2	0.160	0.803	-0.058	0.674
X_3	-0.170	0.753	0.033	0.597
X_4	0.023	0.022	-0.994	0.990
X_5	-0.800	-0.049	-0.990	0.653
X_6	0.808	-0.188	0.087	0.695
X_7	0.956	0.044	0.047	0.918
Variance	3.1753	1.2537	1.0150	5.4440
% Variance	0.454	0.179	0.145	0.778

mendation of Stevens (1992) which recommended that factor loading with absolute values greater than 0.4 which explains about 16% of the variance is meaningful and should be accepted while any variables with factor loading less than 0.4 should be ignored.

From Table 3, factor 1 has a high positive loading for variable x_7 (dry matter content) and variable x_6 carbohydrate and high negative loading for variable x_1 moisture content and x_5 fiber. Factor 1 therefore measures the extent to which fufu flour contains dry matter and carbohydrates rather than moisture and fiber. It can be labeled emphasis on dry matter and a lack of moisture and also emphasis on carbohydrate and lack of fiber. Factor 2 has for variables x_2 and x_3 respectively. Finally, factor 3 has a high negative loading of -0.994 for variable x_4 which is fat.

Conclusion

The factor analysis of the composition of the cassava varieties produced three factors as evidenced by the number of variables whose eigenvalues are greater than one. Factor 1 showed high positive loading indicating high dry matter and carbohydrate content in cassava *fufu* flour. Similarly, protein and ash implicated in factor 2 have high eigenvalues. The proportion of the moisture and fiber on the other hand are low in the *fufu* flour as indicated by their high negative loading.

Finally for factor 3, it was observed that fat is always contained in low proportion in *fufu* flour because of its high negative loading. This study has succeeded in reducing the 43 CMD-resistant varieties to three important groupings. These are Dry matter and carbohydrate in group 1, followed by protein and ash in the next proportion and finally fat in the third proportion.

REFERENCES

Anga SB (2008). Why we need Cassava as a Source of Bio-Fuel. www.businessdayonline.com (Retrieved May 2008).

Crawford IM, Lomas RA (1980). Factory Analysis - a Tool for Data Reduction. Eur. J. Mark. 14(7): 414 – 421.DOI: 10.1108/EUM0000000004917. Emerald Group Publishing Limited

Etudaiye HA, Nwabueze TU, Sanni LO (2009). Quality of *fufu* processed from Cassava Mosaic Disease (CMD) resistant varieties. Afr. J. Food. Sci. ***3***(3): 061-067.

IITA (2005). Growing cassava commercially in Nigeria. Cassava illustration guide book. International Institute of Tropical Agriculture, Ibadan, Nigeria. pp 21 – 22.

Nweke FI, Dunstan S, John L (2002). The Cassava Transformation: Africa's Best Kept Secret. Michigan State University Press, East Lansing. p. 10.

Onyeagu SI (2003). A First Course in Multivariate statistical Analysis. Nnamdi Azikiwe University, Nigeria. Published by MegaConcept, ISBN: 978-2272-29-9.

Pearson K (1927). The Mathematics of Intelligence, The Sampling Errors in the Sampling Theory of a Generalized Factor, Biometrica, 19: 246-292.

Rao PO, Hahn SK (1984). An automated enzymatic assay for determining the cyanide content of Cassava (*Manihot esculenta Crantz*) and Cassava products. J. Sci. Food. Agric. 35: 246- 436

Steven JP (2002). Applied Multivariate Statistics for the Social Sciences (Forth Edition) Published by Lawrence Erlbaum Associates, Hillsdale, N. J. Erlbaim.