Greener Journal of Agricultural Sciences
ISSN: 2276-7770; ICV: 6.15
Vol. 3 (2), pp. 164-175, February 2013

http://gjournals.org/GJAS

Research Article

Study of Some Fruit and Seed Traits Relationship and Assessment of Multicollinearity in Date Palm (*Phoenix Dactylifera* L) Accessions of Nigeria by Correlation and Principal Component Analysis

J.O. Odewale, *Agho Collins, C.D. Ataga, G. Odiowaya, A. Hamza, E.O. Uwadiae and M.J. Ahanon

Plant Breeding Division, Nigerian Institute for Oil Palm Research (NIFOR).P.M.B .1030 , Benin City, Edo State, Nigeria.

ARTICLE INFO

Article No.: *122112341*
DOI: *10.15580/GJAS.2013.2.122112341*

Submitted: *21/12/2012*
Accepted: *10/01/2013*
Published: *20/02/2013*

***Corresponding Author**
Agho Collins
E-mail: *collinsagho@yahoo.com*

Keywords:
date fruit, tolerance value, variance inflation factor, multicollinearity, principal component regression and principal component analysis

ABSTRACT

Collinearity (or multicollinearity) is the undesirable situation where the correlations among the independent variables are strong. Multicollinearity misleadingly inflates the standard errors and results in incorrect conclusions about relationships between dependent and predictor variables. Thus, it makes some variables statistically insignificant while they should be otherwise significant. It is like two or more people singing loudly at the same time. One cannot discern which is which. They offset each other. Most explanatory variables in the biological sciences tend to correlate and this leads to incorrect identification of the most important predictors. Just like other crops, in date palm the seed characters most times are used as predictors of the quality and quantity of the fruits such as the fruit weight and size which are important pricing index of date palm fruits and so, it becomes important to develop a model to explain the relationships between these two data sets without collinearity. The changing environments accompany by floods and high temperature coupled with the increase in population in Nigeria and Africa as a whole is a concern to agriculturist. Famine is fast approaching and crop modelling is one of the solutions. The main purpose of this study is to show how we can use multivariate analysis based on principal component scores to establish a model that can explain the relationship between the fruit and seed traits of date palm and compare its ability to reduce multicollinearity with the method of the ordinary least square while also studying the variability that exist among the germplasm collections for genetic improvement. The result of the descriptive statistics which indicated the values of the coefficient of variation for the different characters of the fruit and seeds of date palm reveals the possibility of genetic improvements of these characters. Principal component analysis (PCA) was applied to predictor variables to address the problem of multicollinearity. The result of the principal component analysis indicated that the contribution of the first two factors with Eigen value greater than unity accounted for 86.5 % of the total variation which was well above average and thus, explains the use of PCA in data reduction. The results showed that the principal component regression (PCR) was sufficient in eliminating multicollinearity with a variance inflation factor (VIF) and tolerance value (TOL) of unity and $P < 0.05$ and it was possible to explain a high percentage of the total variance with a reduced number of principal components as two principal components (PRIN 1 and PRIN 2) accounted for most of the variability in seed characters observed among the date palm germplasm collections from different locations. There was a high level of variation in some of the seed traits studied which could serve as the basis for genetic improvement of date palm fruit. There was a direct positive relationship between the groove width and the fruit traits in the multiple regression analysis and the principle component analysis, thus, indicating the importance of groove width in the genetic improvement of date palm fruits.

INTRODUCTION

Before the advent of scientific knowledge, farmers believed long ago that plants transfer good characters to their offspring. Based on this belief, the more robust and good appearing seeds were selected from farmers harvest and reserved as seed for next season planting (Kortse and Oladiran, 2012). Selection of fruit content is a determinant of fresh fruit quality (Rajan et al., 2005). Awareness of interrelationship between seed and fresh fruit traits of date palm will improve the efficiency of breeding programs for fruit size and other desirable fruit traits through the use of appropriate selection criteria. The seed is usually the first material in the hand of plant breeders because vegetative characters only come to mind after seed planting. Most times knowledge of the interrelationship between seed and fresh fruit traits will help plant breeders to predict the outcome of the fruit that will eventually be formed by visual inspection of the seed characters before planting. When the seed variables are taken as the independent variables and then regressed on the fruit characters (dependent variables), the output will be useful for prediction and selection of desirable qualities of the fruit such as mesocarp thickness, fruit weight, fruit circumference, mesocarp weight which are important index for pricing of date fruit in the market. One important issue in multiple linear regression analysis, and one that seems to be ignored by many biologists who fit multiple regression model to their data, is the impact of the correlated predictor variables on the estimates of parameters and hypothesis tests. If the predictors are correlated, then the data are said to be affected by (multi) collinearity. Severe collinearity can have important and detrimental effects on the estimated regression parameters. Lack of collinearity is very difficult to meet with real biological data where predicted variables that might be incorporated into a multiple regression model are likely to be correlated with each other to a greater extent. The conventional method of prediction in plant breeding is the use of ordinary least square (OLS) which assumed that the variables in the regression equation are independent. The ordinary least square is used to find the best line that on the average is closest to all points and minimises the square residuals. This conventional approach might result in multicollinearity for variables, particularly when correlations among some of the traits are high (Hair et al., 1995). There may be difficulties in interpretation of the actual contribution of each variable (since the effects are mixed or confounded because of collinearity) and supplementation of unique explanatory predictions from additional variables. Most explanatory variables in the biological sciences tend to correlate. It is well known that in the presence of multicollinearity problem, there is a basic violation of one of the assumptions of the ordinary least square estimator and thus the ordinary least squares becomes ineffective as the standard errors of the parameter estimates could be quite high, resulting in unstable estimates of the regression model. Hence, the multicollinearity between predictor variables can lead to incorrect identification of the most important predictors (Sharma, 1996, Thompson et al., 2001; Hoe and Kim, 2004) and this will result in incorrect conclusions about relationships between dependent and predictor variables. Previously, some researchers (Sharma, 1996; Çamdeviren et al., 2005; Sousa et al., 2007; Mendes, 2009) reported that one of the approaches to avoid this problem is the principal component analysis. Recently, the usage of principal component analysis (PCA) to avoid multicollinearity problem began to increase with the availability of related statistical package programs such as SAS, Statistica, SPSS and NCSS (Fievez et al.,2003; Liu et al., 2004; Raick et al., 2006; Posta et al., 2007).An extension of the principal component analysis is the principal component regression (PCR) . The purpose of principal component regression (PCR) is to estimate the values of a response variable at the basis of selected principal components (PCs) of the explanatory variables. There are two main reasons for regressing the response variable on the PCs rather than directly on the explanatory variables. Firstly, the explanatory variables are often highly correlated (multicollinearity) which may cause inaccurate estimations of the least squares (LS) regression coefficients. This can be avoided by using the PCs in place of the original variables since the PCs are uncorrelated. Secondly, the dimensionality of the regressors is reduced by taking only a subset of PCs for prediction. Plant breeders are seldom interested in a single trait and therefore, there is the need to examine the relationships among various traits, especially between seed and fruit traits. As the number of independent variables influencing a particular dependent variable is increased, a certain amount of interdependence is expected. In such situations, correlations may be insufficient to explain the associations in a way that will enable breeders to decide on a direct or indirect selection strategy (Ofori, 1996).The quality of the seed determines the nature of the fruit that will eventually be formed at maturity since the fruit is usually produced from the seed. During selection of seeds for planting, plant breeder normally concentrate on the desired traits of the seed because they are aware of the interrelationships between the traits of the seed and the fruit that will eventually be formed. Knowledge of an accurate method of estimating the interrelationships between the traits of these two data sets will be necessary for the genetic improvement of the fruit such as large fruit size, high mesocarp weight and thick mesocarp. Multiple regression analysis has been used to interpret the complex relationships in biological variables. However, its interpretation may be misleading where there exists multicollinearity among the predictor variables. To address this limitation, a multivariate analysis such as the principal component analysis is more suitable as a statistical method for reducing a complex system of correlations into one of smaller dimensions through the extraction of a few unobservable latent variables called factors (Tabachnick and Fidell, 2001). Factor scores

can be derived from such multivariate analysis which could be nearly uncorrelated or orthogonal. Such factor scores could therefore be used for prediction, thereby solving the problem of multicollinearity which is vital in crop modelling and genetic improvement. There is need to increase date palm fruit productions in Nigeria because according to FAO, 1999, Nigeria date palm production figures are usually not reflected in FAO production. In Nigeria, sub-Saharan Africa, information on the use of multivariate statistical approach to elucidate the structural relationships among morphometric traits of date palm is insufficient. Therefore, the present investigation is aimed at: (i) Studying the relationship between seed and fruit traits in Nigerian date palm so as to ascertain the existence or otherwise of collinearity instability which is usually ignored by plant scientist. If detected, the second objective was to use a multivariate approach, in this case, principal component regression analyses to try and address the problem. (ii) Estimating the fruit characters of date palm (Phoenix dactylifera l.) from the measurements of the seed characters based on principal component regression. (iii) Estimating the genetic variability in the date palm accessions for genetic improvements.

MATERIALS AND METHODS

Data collection and analysis: The main date palm-growing areas of Nigeria were surveyed in 2011 with the objective of characterizing cultivars as to the quality and economic value of their fruits. Ten new accessions were collected and measured, all under cultivation and in full production in the farmers' field. 100 fruits were randomly selected from different palms of each of the accessions and mixed together to act as replicates, and then 20 fruits were taken randomly for each accession. For the (20) fruits, the following traits were measured: fruit length (FLT), weight (FW), mesocarp thickness(MT), mesocarp weight(MW),seed weight(SW),seed length(SL),seed circumference(SC),groove length(GL),groove width(GW). The measurements for each dimension were replicated 10 times. Mass of individual fruit was determined using an electronic balance with a sensitivity of 0.0l g while the length, thickness and circumference were determined using a vernier calliper with a reading accuracy within 0.01 mm.

Test For multicollinearity among the preditors.

a. Examining the correlation matrix of the predictor variables.

First, we should examine a matrix of correlation between the predictor variables, and look for large correlations .A correlation greater than 0.7 is an indication of the presence of collinearity.

b. Variance inflation factor (VIF) and tolerance value (TOL).

Marquardt (1970) gives the diagonal elements of the matrix R-1xx = (X'X)-1 of variance inflation factor (VIF) when the matrix X'X is taken in the correlation form. This factor can be used to detect multicollinearity (Montgomery and Peck, 1981).According to Neter et al. (1983),if VIF has a value greater than 10, it is possible that the minimum squares regression coefficients associated with such values are highly affected by multicollinearity. Any VIF greater than 10.0 indicates strong collinearity in the data set. Secondly, we should check the tolerance value for each predictor variable. A low tolerance indicates that the predictor variables is correlated with one or more of the other predictors and a TOL less than 0.10 is enough to worry about.

c. Condition number and Eigen value in the collinearity diagnostic.

Condition Numbers: where and were the maximum and minimum eigenvalues respectively. If the condition number of the correlation matrix exceeded 1000, or exceeded 30, the effect of multicollinearity should be considered. Also, principal component with Eigen value nearer to zero indicates collinearity between the original predictor variables, because those components have little variability that is independent of the other component. When there is no collinearity at all, the eigenvalues, condition indices and condition number will all equal one. As collinearity increases, eigenvalues will be both greater and smaller than 1 (eigenvalues close to zero indicate a multicollinearity problem).

Multiple Linear Regression Analysis (MLR)

MLR include a large number of independent (predictor) variables (X's) where some of them may be redundant because of high correlations (multicollinearity problem) with other independent variables. The use of redundant predictors can be harmful since potential gain in accuracy attributable to their inclusion is outweighed by inaccuracies associated with estimating their proper contribution to the prediction (Spark et al., 1985). Especially, in the presence of multicollinearity among the column of X can have significant impacts on the quality and stability of the fitted model (Johnson, 1991). One approach to avoid this problem is PCA (Mendes, 2009). This is achieved by transforming set of original variables to a new set of variables, the principal components (PCs), and which are ordered so that the first few retain most of the variation present in all of the original variables (Joliffe 2002).Since detailed information about PCA was given by Johnson and Wichern (1982), Sharma (1996) and Tabachnick and Fidell (2001), we did not give any further information about PCA. The basic equation of PCA is, in matrix notation, given by following equation:

$$Y=W'X$$

Where X is the matrix which contains p original variables and W is the matrix of weights which contains standardized weights (wij) of each variable in PCs. The magnitudes of the coefficients give the

contributions of each variable to that component. The units of the variables could have an effect on PCA. Due to differences in the units of variables used in PCA, correlation matrix of variables (C) was used to get eigenvalues and weight of variables. Where p is the number of components, n is the number of observations in the sample and C is the correlation matrix. The varimax criterion of the orthogonal rotation method was employed in the rotation of the factor matrix to enhance the interpretability of the principal components. Bartlett's sphericity test was used to test the null hypothesis that the correlation matrix is an identity matrix or that all correlations are zero. In other words, the applicability of PCA is determined based on Bartlett's sphericity test result. Therefore, the applicability of PCA was tested by Bartlett's sphericity test. The formula for computing this test is:

$$X^2 = \frac{(11 + 2p - 6n) * \log_e|C|}{6}$$

Where p is the number of components, n is the number of observations in the sample and C is the correlation matrix.

In this study, we used only two PCs with eigenvalue greater than 1 out of 5 PCs (Kaiser 1960). Therefore, only score values of first two selected PCs were considered. Coefficient of determination (R2), Durbin-Watson statistic (DW) and residual mean square error (RMSE) were used as goodness-of-fit criteria. SPSS version 17.0 statistical package programs were used in the statistical analyses.

RESULT

A method for diagnosing multicollinearity should supply information about the degree of manifestation, and further identify the variables involved in the problem. A closer look at the size of the non-diagonal elements of the matrix (Table 1) using the seed traits (X6-X10) shows a correlation values that approaches unity (0.826 and 0.899) which were highly significant (P< 0.001) showing a high degree of multicollinearity among some of the seed characters measured. To further confirm this hypothesis, the tolerance value (TOL) and variance inflation factor (VIF) were measured. When the raw data of the seed traits were used for the multiple linear regression analysis (MLR), a multicollinearity problem was detected with a high values of variance inflation factor (VIF>10.0; Table 5). The tolerance value which is an inverse of the VIF also confirms the presence of multicollinearity in the seed characters studied with some values very close to zero (Table 5) when the ordinary least square (OLS) method was used. The weight and length of seeds in date palm seems to have strong multicollinearity problem with some other seed characters studied. The values of the coefficient of determination for the regression of the seed characters on each of the fruit traits in the MLR ranges from 73.6 % to 97.0 % with groove width having the highest regression coefficient in the MLR.The circumference of the seed were found to have a negative influence on all the fruit characters studied with a negative regression coefficient in all the fruit traits. Bartlett's sphericity test for testing all correlations are zero or for testing the null hypothesis that the correlation matrix is an identity matrix was used to verify the applicability of PCA. The value of Bartlett's sphericity test was found to be 31.54 and it implied that the PCA is applicable to our data set (P=0.000). In Figure 1(A biplot of the component scores), two clusters were identified which were closely associated together (presence of multicollinearity).Seed weight, circumference and width were closely associated together while groove length and seed length were also closely associated. Two principal components were extracted from the seed traits studied. These two components (rotated with varimax method) together accounted for 86.5% of the total variation of the variables in the principal component analysis. First component (52.9%) strongly influenced seed weight, circumference and groove width while the second component (33.6%) influenced mostly seed length and groove length. Communality values of variables (Table 4) were also found to be high in weight of the seed (83 %), length of the seed (96.3 %), circumference of the seed (94.5 %), groove length (93.6%), and groove width (65.2 %). The highest values of communalities indicate that the variances of variables were efficiently reflected by variables in MLR analysis. Therefore, only the first two principal components were appropriate for explaining the variations in date palm seed traits studied. A condition indices test based on the Eigen value in the collinearity diagnostic was conducted and a value of 159.9 was obtained which was far higher than 30(a standard value), therefore, confirming the presence of multicollinearity in the seed characters studied. To address the problem of multicollinearity, the principal component regression method was used before and after stepwise regression. Stepwise regression analysis was made to determine which PCs contributed to the variation in the dependent (fruit characters) variables set. When the mesocarp thickness was used as dependent variable, stepwise regression analysis revealed that only PC2, which was composed of seed length and groove length, contributed the most in the prediction of mesocarp thickness and it was significant (P < 0.05). The Variance Inflation Factor (VIF) and the Tolerance values (1.00 in both cases) indicated that the problem of multicollinearity has been addressed (Table 2). The basic descriptive statistics [mean, standard deviation, standard error and coefficient of variation (CV)] of seed and fruit traits of date palm are presented in Table 6. A high value of coefficient of variation was obtained for seed weight, groove length; groove width and mesocarp thickness. The CV indicates the level of

variation of the traits which could serve as a basis for genetic improvement of date palm fruit. When all the characters (fruit and seed traits) were used as variables in the PCA (Table7), three components with Eigen value greater than unity were extracted and accounted for 97.03% of the variability. The first component which accounted for 60.2% of the variability were primarily loaded with all the fruit characters and Seed groove width, with fruit weight having the highest loading (0.907) while the second component was primarily loaded with seed length and groove length. The third component was loaded with seed weight and circumference (Table 8).

DISCUSSION

The result of the descriptive statistics which indicated the values of the coefficient of variation for the different characters of the fruit and seeds of date palm reveals the possibility of genetic improvements of these characters as the levels of variability of some of the characters namely seed weight, groove length, groove width and mesocarp thickness were high. The Pearson-moment correlations reveal high correlation coefficients between the seed and fruit characters in general. Therefore, these variables were used to predict each of the fruit traits which include mesocarp weight, mesocarp thickness, fruit length and fruit weight. Generally, high correlations were also found between seed traits (predictor variables). These correlations provide a measure of the linear relations between two variables and also indicate the existence of multicollinearity problem between the predictor variables (Sharma, 1996).When the raw data of the seed characters were used for the multiple linear regression, multicollinearity with high value (VIF > 10) existed and thus the use of interdependent explanatory variables should be treated with caution since multicollinearity has been shown to be associated with unstable estimates of regression coefficients (Malau-Aduli et al., 2004), rendering the estimation of the unique effects of the predictors impossible. Another alarming point is that some of the individual t-ratios of the coefficients were non-significant which again is an indication of the presence of multicollinearity among the predictors. Looking at the tolerance and VIF columns, only one of the independent variables (groove width) is causing less to multicollinearity, but the rest show a presence of high correlation. This justifies the use of principal component factor scores for prediction. These factors are orthogonal to each other and are more reliable in weight estimation. PC1 and PC2 together accounted for 86.5% of the variation in seed characters of date palm. Only the first two principal component axes (PC1 and PC2) in the PCA analysis had Eigen values up to 1.0, presenting cumulative variance of 86.5%. Principal component one (PC1), with Eigen value of 2.647, contributed 52.9% of the total variability, while PC2, with Eigen value of 1.679, accounted for 33.6% of total variability observed among the 10 date palm genotypes. In PC1, the traits that accounted for most of the 52.9% observed variability among the 10 genotypes included weight of the seed, circumference of the seed and groove width while PC2 were primarily loaded with the length of the seed and groove length. The Variance Inflation Factor (VIF) and the Tolerance values (1.00 in both cases) obtained when the PC scores were used for the regression analysis indicated the complete removal of multicollinearity. The high values of the communalities obtained from the PCA confirm the effectiveness of using only the two retained factors to explain the variations of these variables in the multiple regressions. Therefore, MLR model based on principal component scores for investigating the relations between seed and fruit traits (using mesocarp thickness as an example) can be written as:

Mesocarp thickness = 0.002PC1 + 0.063PC2.

Stepwise regression analysis was made to determine which PCs contributed to the variation in the dependent (fruit characters) variables set. The stepwise regression analysis revealed that only PC2, which was composed of length of the seed and groove length, contributed the most and the prediction was significant ($P < 0.05$). Hence, the final MLR model can be written as:

Mesocarp thickness = 0.063PC2

All other dependent variables such as fruit length, mesocarp weight and fruit circumference also had a VIF and tolerance value of unity when regressed on the PC scores obtained from the seed characters (Tables not shown) and thus their MLR model can also be determined accurately without multicollinearity. When the seed characters were regressed on each of the fruit traits, the groove width was found to have the highest positive contribution to each of the fruit traits in the multiple regression thus, indicating the direct positive relationship between the groove width and the fruit traits and thus the higher the groove width the higher the fruit components and this could be used for genetic improvement of the fruit components by date palm breeders. In the principal component analysis of the fruit and seed characters, only the groove width of the seed among the seed characters studied were loaded along with all the fruits characters in the first principal components. The inclusion of the groove width, a high positive loading, along with the fruit characters in the first component further confirms the fact that the selection for a large groove width of the seed in date palm could be used for the genetic improvement of date palm fruit .The highest loadings of all the fruit traits along the first principal components when both the seed and fruit traits were used in the PCA suggested that the fruit characters are sufficient in separating date palm accessions. It has been observed that when the raw data were used for the regression analysis, the multicollinearity problems existed (VIF > = 10.0). On the other hand, when the transformed data in the form of PC scores were included in the multiple regression analysis as predictor variables instead of original predictor values, multicollinearity disappeared (VIF = 1) and it was possible to explain a high percentage of the total

variance explained with a reduced number of principal components as two principal components (PRIN 1) and PRIN 2) accounted for most of the variability observed among the date palm germplasm collections from different locations. Similar results were obtained by Maji and Shaibu (2012), working on rice germplasm evaluation. Therefore, using the principal component scores in multiple regression analysis for predicting fruit traits in date palm is more appropriate than using the original seed traits data. As a result it can be stated that the use of principal component based model is considered as more efficient, due to elimination of multicollinearity problem, reduction of the number of predictor variables, decrease of the model complexity and better interpretation of multiple regression models by removing indirect effects related to predictor variables. This study has contributed significantly towards finding out lead variables in date palm and also towards reduction of multiplicity of variables and concentrating on the lead variables for date palm improvement program. This method has been used for data reduction, grouping of variables and also to find out the lead variables in different crops like cardamom (Radhadrishan et al., 2004); coconut (Odewale et al., 2012), peanut (Ajay, 2012), oil palm (Ataga, 2009).Egypt is the highest producer of date palm in Africa. Algeria and Sudan are the second and third highest in Africa respectively (FAO, 1999).Nigeria date palm production is not reflected in FAO production year books (NIFOR 37th Annual Reports, 2000) and a knowledge of the variability in the germplasm collection as revealed by the principle component analysis coupled with accurate modelling will help date palm breeders in Nigeria in the genetic improvement of date palm fruits which consequently enhances production.

Table 1a: Simple correlation coefficients among seed traits measured in date palm genotypes.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1	1	0.965**	0.963**	0.834**	0.483	0.554	0.303	0.489	0.737*	0.804**
X2	0.804**	1	0.740*	0.749*	0.782**	0.368	0.684*	-0.010	0.624	0.307
X3	0.965**	0.740*	1	0.938**	0.815**	0.520	.506	0.357	0.494	0.767**
X4	0.963**	.749*	0.938**	1	0.798**	0.449	0.558	0.317	0.496	0.758*
X5	0.834**	0.782**	0.815**	0.798**	1	0.099	0.600	-0.198	0.650*	0.424
X6	0.483	0.368	0.520	0.449	0.099	1	0.481	0.826**	0.279	0.494
X7	0.554	0.684*	0.506	0.558	0.600	0.481	1	0.068	0.899**	0.148
X8	0.303	-0.010	0.357	0.317	-0.198	0.826**	0.068	1	-0.032	0.676*
X9	0.489	0.624	0.494	0.496	0.650*	0.279	0.899**	-0.032	1	0.203
X10	0.737*	0.307	0.767**	0.758*	0.424	0.494	0.148	0.676*	0.203	1

Table 1b: Description of the variables

Variables	Description
X1	FRUITWEIGHT
X2	FRUITLENGHT
X3	FRUITCIRCUMFERENCE
X4	MESOCARPWEIGHT
X5	MESOTHICKNESS
X6	SEEDWEIGHT
X7	SEEDLENGHT
X8	SEEDCIRCUMFERENCE
X9	GROOVELENGHT
X10	GROOVEWIDHT

Table 2: Multivariate multiple linear regression analysis results based on PCs scores

Variable	DF	Estimates	Standard Error	T value	P	R^2	VIF	TOL
Before stepwise regression analysis								
PC1	1	0.002	0.028	0.085	0.934	41.4	1.00	1.00
PC2	1	0.063	0.028	2.221	0.062		1.00	1.00
After stepwise regression analysis								
PC2	1	0.063	0.027	2.373	0.045	41.3	1.00	1.00

Table 3: Eigenvalues, individual and cumulative % of seed characters.

Principal components	*Eigenvalue*	*Individual(%)*	*Cumulative(%)*
1	*2.647*	52.943	52.943
2	1.679	33.586	86.529
3	0.548	10.967	97.496
4	0.089	1.787	99.283
5	0.036	0.717	100.000

Table 4: Factor scores of characters associated with the first two principal axes of ordination along with final communalities of date palm seed.

Character	Axis of Ordination		Communality
	I	II	
Weight	**0.850**	**0.328**	**0.830**
Length	**0.145**	**0.971**	**0.963**
Circumference	**0.969**	**-0.079**	**0.945**
Groove length	**0.045**	**0.966**	**0.936**
Groove width	**0.803**	**0.084**	**0.652**

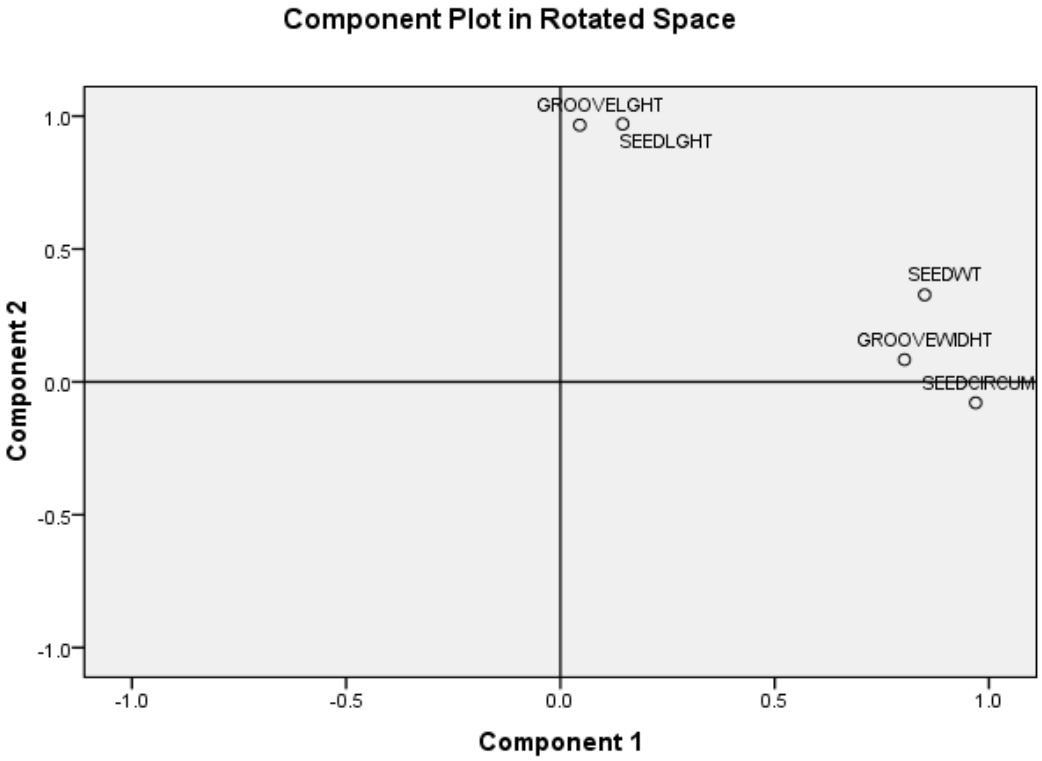

Figure 1: Component plot of seed traits

Table 5: Multivariate multiple linear regression analyses results based on raw data.

FRUITWEIGHT

Variable	Estimates	Std. Error	T value	P	VIF	TOL
Constant	18.494	16.850	1.098	0.334		
Weight	6.256	4.131	1.514	0.205	10.168	0.098
Length	8.188	7.241	1.131	0.321	11.893	0.084
Circumference	-12.165	5.16	-2.511	0.066	9.509	0.105
Groove length	-6.274	5.378	-1.161	0.308	7.868	0.127
Groove width	69.338	13.173	5.264	0.066	2.430	0.411
R^2=92.2 %, STD=1.341, MODEL P=0.025						

FRUITLENGHT

Variable	Estimates	Std. Error	T value	P	VIF	TOL
Constant	6.981	4.377	1.595	0.186		
Weight	1.514	1.073	1.410	0.231	10.168	0.098
Length	0.231	1.881	0.123	0.908	11.893	0.084
Circumference	-2.430	1.341	0.123	0.144	9.509	0.105
Groove length	0.036	1.397	0.026	0.981	7.868	0.127
Groove width	5.872	3.422	1.716	0.161	2.430	0.411

R^2=73.6 %, STD=0.384, MODEL *P*=0.229

FRUIT CIRCUMFERENCE

Variable	Estimates	Std. Error	T value	P	VIF	TOL
Constant	12.665	5.600	2.262	0.087		
Weight	2.804	1.373	2.042	0.111	10.168	0.098
Length	-0.356	2.407	-0.148	0.889	11.893	0.084
Circumference	-4.339	1.716	-2.529	0.065	9.509	0.105
Groove length	0.091	1.787	0.051	0.962	7.868	0.127
Groove width	19.747	4.378	4.511	0.011	2.430	0.411

R^2=90.30 %, STD=0.446, MODEL *P*=0.037

MESOCARP WEIGHT

Variable	Estimates	Std. Error	T value	P	VIF	TOL
Constant	2.615	18.294	0.143	0.893		
Weight	2.301	4.485	0.513	0.635	10.168	0.098
Length	13.254	7.862	1.686	0.167	11.893	0.084
Circumference	-8.808	5.606	-1.571	0.191	9.509	0.105
Groove length	-8.348	5.839	-1.430	0.226	7.868	0.127
Groove width	69.354	14.302	4.849	0.008	2.430	0.411

R^2=91.3 %, STD=1.456, MODEL *P*=0.030

MESOCARP THICKNESS

Variable	Estimates	Std. Error	T value	P	VIF	TOL
Constant	1.294	0.321	4.038	0.016		
Weight	0.259	0.079	3.294	0.030	10.168	0.098
Length	0.022	0.138	0.158	0.882	11.893	0.084
Circumference	-0.622	0.098	-6.328	0.003	9.509	0.105
Groove length	0.032	0.102	0.312	0.770	7.868	0.127
Groove width	2.031	0.251	8.106	0.001	2.430	0.411

R^2=97.0 %, STD=0.0255, MODEL *P*=0.0040

Table 6: Descriptive statistics for seed and fruit characters of date palm

	Mean	Std. Deviation	Standard error	Coefficient of variation
SEEDWEIGHT	1.634	0.345	1.011	12.55
SEEDLENGHT	2.591	0.213	0.143	3.85
SEEDCIRCUM	3.029	0.267	0.302	4.62
GROOVELGHT	2.183	0.234	1.040	15.06
GROOVEWIDHT	0.160	0.054	0.031	10.84
FRUITWT	8.060	3.20	0.109	6.68
FRUITLGHT	3.711	0.452	0.067	2.60
FRUITCIRCM	6.537	0.955	0.084	2.79
MESOWEIGHT	6.908	3.29	0.074	3.38
MESOTHICKNS	0.286	0.098	0.017	10.46

Table 7: Eigen values, individual and cumulative % of seed and fruit characters.

Principal components	*Eigenvalue*	*Individual,%*	*Cumulative, %*
1	6.020	60.196	60.196
2	2.100	20.999	81.195
3	1.166	11.662	92.857
4	0.417	4.173	97.029

Table 8: Factor Scores of Characters Associated with the First Two Principal Axes of Ordination along with final communalities of date palm seed and fruit traits.

Character	Axis of Ordination			Communality
	I	II	III	
Seed Weight	0.172	0.345	0.882	0.926
Seed Length	0.229	0.943	0.152	0.965
Seed Circumference	0.139	-0.117	0.979	0.991
Seed Groove length	0.264	0.889	0.010	0.860
Seed Groove width	0.756	-0.129	0.549	0.890
Mesocarp weight	0.894	0.308	0.233	0.962
Mesocarp thickness	0.817	0.479	-0.253	0.977
Fruit circumference	0.896	0.290	0.277	0.802
Fruit length	0.641	0.625	-0.02	0.963
Fruit weight	0.907	0.324	0.224	0.949

REFERENCES

Ataga, CD (2009). Genetic diversity among Nigerian collections of oil palm Elaeis guineensis jacq as revealed by principal component analysis and minimum spanning tree. Journal of Agriculture, forestry and fisheries (volume 10).

Fievez V, Vlaeminck B, Dhanoa MS, Dewhurst RJ (2003). Use of principal component analysis to investigate the origin of heptadecenoic and conjugated linoleic acids in milk. Journal of Dairy Science. 86: 4047- 4053.

Hair JR, Anderson RE, Tatham RL, Black WC (1995). Multivariate data analysis with readings. Prentice Hall, Englewood, NJ.

Hoe JS, Kim DS (2004). A new method of ozone forecasting using fuzzy expert and neural network systems. Science of the Total Environment. 325:221-237.

Johnson JD (1991). Applied Multivariate Data Analysis. Springer-Verlag New York, USA.

Johnson RA, Wichern DW (1982). Applied Multivariate Statistical Analysis (5th edition). Upper Saddle River, NJ:Prentice Hall.

Joliffe I (2002). Principal Component Analysis, 2nd ed., Springer.

Kaiser HF (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement. 20:141-151.

Kortse P Aloho, Oladiran A Johnson (2012). Effects of fruit size on the quality of 'Egusi-itoo' melon (Cucumeropsis mannii Naudin) seed. Advances in Applied Science Research. 3 (4):2192-2195.

Liu Y, Lyon BG, Windham WR, Lyon CE, Savage EM (2004). Principal component analysis of physical, color and sensory characteristic of chicken breast deboned at two, four, six and twenty-four hours postmorten. Poultry Science. 83: 101-108.

Maji AT, Shaibu AA (2012). Application of principal component analysis for rice germplasm characterization and evaluation. Journal of Plant Breeding and Crop Science Vol. 4(6): 87-93.

Malau-Aduli AEO, Aziz MA, Kojina T, Niibayashi T, Oshima K, Komatsu M (2004). Fixing collinearity instability using principal component and ridge regression analyses in the relationship between body measurements and body weight in Japanese Black cattle. Journal of Animal and Veterinary Advances. 3: 856–863.

Marquardt,D.W.(1970).Generalized inverse,ridge regression,biased estimation, and nonlinear estimation.Technometrics 12: 591- 612.

Mendes M (2009). Multiple linear regression models based on principal component scores to predict slaughter weight of broilers. Archive Gefligelkünde. 73(2):139-144.

Montgomery,D.C. and Peck, E.A.(1981).Introduction to Linear Regression Analysis.John Wiley and Sons,New York,pp.504.

Neter J,WassermanW, Kutner MH (1983). Applied linear regression models. Richard D. Irwin Inc., Homewood, pp.547.

Ofori I (1996). Correlation and path-coefficient analysis of components of seed yield in Bambara groundnut (Vigna subterranean). Euphytica .91: 103-107.

Posta J, Komlosi I, Mihok S (2007). Principal component analysis of performance test traits in Hungarian Sport horse mares. Archives of Animal Breeding. 50: 125-135.

Radhadrishan VV, Priya P, Menon KJ, Madhusoodanan KM, Kuruvilla, Thomas (2004). Factor analysis in cardamom (Elettaria cardamomum Maton) J. Spices Aromatic Crops.13:37-39.

Raick C, Beckers JM, Soetaert K, Gregoire M (2006).Can principal component analysis be used to predict the dynamics of a strongly non-linear marine biogeochemical model? Ecological Modelling. 196: 345-364.

Rajan S, Yadava LP, Ram Kumar, Saxena SK (2005). Selection possibilities for seed content: A determinant of fresh fruit quality in guava (Psidium guajava L.). Journal of Applied Horticulture. 7(1):52-54.

Sharma S (1996). Applied multivariate techniques. JohnWiley & Sons, Inc., Canada.

Sousa S IV, Martins FG, Alvim-Ferraz MCM, Pereira MC (2007). Multiple linear regression and artificial neural Networks based on principal components to predict ozone concentrations. Environmental Modelling and Software. 22: 97-103.

Spark RS, Zucchini W, Coutsourides D (1985). On variable selection in multivariate regression. Communications in Statistics - Theory and Methods. 14(7):1569-1587.

Tabachnick BG and Fidell LS (2001). Principal components and factor analysis. In BG Tabachnick & LS Fidell, Using multivariate statistics. (4th ed.). Needham Heights, MA, Allyn & Bacon. pp. 582 - 633.

Tabachnick BG and Fidell LS (2001). Using Multivariate Statistics. Allyn and Bacon A Pearson Education Company Boston, U.S.A.

Thompson ML, Reynolds J, Cox LH, Guttorp P, Sampson PD (2001). A review of statistical methods for the meteorological adjustment of tropospheric ozone. Atmospheric Environment. 35: 617-630.

Cite this Article: Odewale JO, Agho C, Ataga CD, Odiowaya G, Hamza A, Uwadiae EO, Ahanon MJ, (2013). Study of Some Fruit and Seed Traits Relationship and Assessment of Multicollinearity in Date Palm (*Phoenix dactylifera* L) Accessions of Nigeria by Correlation and Principal Component Analysis. Greener Journal of Agricultural Sciences. 3(2):164-175, http://doi.org/10.15580/GJAS.2013.2.122112341**.**