Science, Technology and Development 36 (3): 142-146, 2017
ISSN 0254-6418 / DOI: 10.3923/std.2017.142.146

Exploring Relationship among Quantitative Traits of Sugarcane Varieties Using Principal Component Analysis

[1]Irum Raza, [2]Muhammad Asad Farooq, [1]Muhammad Asif Masood, [1]Saleem Abid, [1]Muhammad Zubair Anwar, [1]Masooma Hassan and [2]Ruqeah Mustafa

[1]Social Sciences Research Institute, National Agricultural Research Center, Islamabad, Pakistan
[2]Crop Science Institute, National Agricultural Research Center, Islamabad, Pakistan
[3]Department of Mathematics and Statistics, University of Agriculture, Faisalabad, Pakistan

Abstract: This study was designed to trace relationship among the stalk and yield characteristics of fifteen sugarcane varieties, using principal component analysis. Ten quantitative traits, such as, stalk height (cm), stalk diameter (cm), single cane weight, number of internodes per stalk, internodal length, leaf area index, number of tillers per plant, cane yield (tons ha^{-1}), commercial cane sugar and total biomass were selected for data analysis. Principal component analysis categorized ten traits into four groups and eigen values greater than one were used to assign principal components into PC1, PC2, PC3 and PC4. These components explained 84.5% of the total variance within the varieties. Single cane weight, leaf area index, stalk height (SH), intermodal length (IL) total biomass and internodes per stalk had direct and valuable impact on sugarcane yield in the first three components, and these characters could be useful and desirable tools for improvement in sugarcane yield.

Key words: Quantitative traits, Principal component analysis, Sugarcane.

INTRODUCTION

Sugarcane crop is among one of the most important salable crops of Pakistan, which is cultured for sugar and sugar related production. Sugarcane contributes a share of 3.1 and 0.6% in value addition in agriculture and GDP, respectively (GoP, 2014). In sugarcane breeding, selection of appropriate varietal indices is extremely important for improving seed yield quality (Bahmankar *et al.,* 2014). Evaluating genetic relationships amid stalks and yield traits is essential for improvement and breeding in sugarcane. Relationships among phenological and morphological traits of crops, such as, sugarcane and safflower, have been found with the computation of principal component analysis (PCA) technique (You *et al*., 2013); Mozzafari and Asad, 2006); Ahmadzadeh *et al.,* 2010).

In varietal trials, plant breeders are usually interested in finding the most suitable variety based on the multiple characteristics. For this purpose, they make separate analysis to study different characters of a variety but this approach does not lead to meaningful statistical inference (Singh and Singh, 2013). The multivariate techniques, such as, PCA, factor analysis, discrimination and classification, cluster and correspondence analysis, can be applied to study multiple characters simultaneously (Khattree and Naik, 2000).

Genetic diversity in different crops has been analyzed through Multivariate statistics; barley by Cross (1992), sorghum by Ayana and Bekele (1999), wheat (*Triticum* sp.) by Hailu *et al.* (2006), peanut by Upadhyaya *et al.* (2009) and vineyard peach by Nikolic *et al*. (2010) and in rice Bharadwaj *et al.*, (2001). A parallel research was conducted, using PCA, to evaluate the correlation between morphology, phenology and yield behaviour of 20 dissimilar safflower genotypes (Bahmankar *et al.,* 2014).

Traditionally different statistical techniques, such as, regression and correlation, have been used to study the characteristics of crop varieties separately. An advantage of PCA over other statistical techniques is that it reduces data in two dimensions, i.e., it transmutes a set of linearly correlated independent variables into a set of uncorrelated variables (Raychaudhuri *et al.,* 2000; Smith, 2002).

Keeping in view the importance of PCA in agriculture, the present study was planned in collaboration with sugarcane coordinated program of National Agricultural Research Centre (NARC). In the current study focus was made on studying all of the factors that affect the cane yield of sugarcane simultaneously. The main goal of the study was to find relationship between stalk and yield characters of sugarcane crop varieties using PCA.

Corresponding Author: Irum Raza, Social Sciences Research Institute, National Agricultural Research Center, Islamabad, Pakistan
E-mail: irumssrinarc@gmail.com

MATERIALS AND METHODS

Present study has been designed in collaboration with sugarcane coordinated program of NARC. Data on stalk and yield characters of fifteen sugarcane varieties included stalk height (cm), stalk diameter (cm), single cane weight, internodes per stalk, internodal length, leaf area index, number of tillers per plant, cane yield (tons ha^{-1}), commercial cane sugar and total biomass.

In agricultural field experiments, the data on multiple characters are observed. For instance, if we have many variables affecting the crop of variety then we do not consider all the factors at a time because it does not draw statistical inference. Instead a separate analysis is made for each variable and interpretation is made accordingly. In the present study, focus is made on analyzing the effect of all of the variables on the response simultaneously. This target has been achieved by means of using multivariate statistical procedure named PCA. Several advantages of this technique have been highlighted in many research studies, such as, Singh and Singh (2013), Micó *et al.* (2006) and Rotaru *et al.* (2013).

The main goal of PCA is to explain the maximum variance through a few number of principal components. PCA has many applications in agriculture, social science, marketplace research and other industries, where experiments are based on a multitude of variables. We can use PCA to reduce the number of variables and avoid multi-colinearity as stated by Jolliffe (2002).

Shlens (2014) has proposed the computation of PCA in the following steps:

1. Arrange data as an m×n matrix, where m represents the number of measurement types and n symbolizes number of samples.
2. Subtract off the mean for each measurement type.
3. Compute first the correlation matrix and then eigen values of the correlation matrix.

Deriving principal components: Derivation of principal components prescribed by Jolliffe (2002) is given by

$$\mathrm{Var}\{\alpha_1 x\} = \alpha_1 \sum \alpha_1$$

The constraint used in the derivation is $\alpha 1.\alpha 1 = 1$, that is, the sum of squares of elements of $\alpha 1$ equals one.

In general, the *k*th PC of x is a'kx and Var $\{\alpha_k x\}$ = □

Data were subjected to MINITAB software version 16 and principal component analysis was performed.

Eigen values, variance percentage and cumulative percentage were found. Scree plot and score plot were also obtained in order to decide how many principal components are sufficient to describe the relationship.

RESULTS AND DISCUSSION

First of all, analysis of variance was performed to assess the significance of individual variables. The mean square errors in Table 1 show variation among different varieties. This clearly indicates that these variables have an effect on the yield of different varieties of sugarcane.

Table 1: Mean squares of the analysis of variance by stalk and yield characters of fifteen sugarcane varieties.

s.o.v	df	SH	SD	SCW	NIS	IL	LAI	NTP	CY	CCS	TB
replication	2	228.71	1.58	0.00	0.87	0.77	0.07	0.10	3.37	0.30	31.88
varieties	14	547.89*	28.47**	0.05**	5.83**	5.01***	1.26**	2.58**	963.70**	14.63**	20.32**
residuals	28	265.82	2.38	0.00	0.95	2.54	0.23	0.09	13.32	0.59	3.52

SH: Stalk height (cm); SD: Stalk diameter (cm); SCW: Single cane weight; NIS: No. of internodes per stalk; IL: Internodal length; LAI: Leaf area Index; NTP: No of tillers per plant; CY: Cane yield (tons ha-1); CCS: Commercial cane sugar; TB: Total biomass
***=significant at 5% alpha, **= significant at 1% alpha, ***= significant at 10% alpha by F-test**

Next step was to obtain the correlation matrix among all of the fourteen variables as depicted in Table 2. Figures, in bold, show positively and significantly correlated variables that could be further analyzed, using PCA to derive the key variables that have effect on the yield of sugarcane varieties. It is apparent from Table 2 that the variables stalk diameter, single cane weight, no. of internodes per stalk and leaf area index have correlation values close to 1 and are positively correlated with sugarcane yield.

Eigen values of the correlation matrix explain the partitioning of the total variance accounted for each principal component. Variance percentage explained by each Eigen value and the cumulative proportion of the variance for all the components are given in Table 3. Eigen values greater than one was taken to decide the number of principal components as suggested by Panagiotakos *et al*., (2007). In this case, eigen values corresponding to the first four variables were considered because they all had eigen values greater than one. Similarly contribution of variance and cumulative percentage were decided accordingly.

PCA was applied to evaluate relationship among stalk and yield characters of fifteen sugarcane varieties. PCA placed ten characters into four groups and eigen values greater than one was used to allocate components into PC1, PC2, PC3 and PC4 as given in Table 4. These components explained 84.5% of the total variance. PC1 justified highest variance (35.5%). The positive values of components corresponding to the characters single cane weight (SCW), leaf area index (LAI) and cane yield (CY) were 0.418, 0.477 and 0.491 respectively. PC1 was called a basic

component of seed yield due to high and positive values in this component. PC2 had explained 21.5% of the variance and showed the positive values for the characters stalk height (SH), internodal length (IL) and total biomass and named as stalk height and internodal length component. PC3 had shown 17.3% of the variance and only one positive value for the character number of internodes per stalk was found so this component was named internodes per stalk component. PC4 was capable of explaining only 10% of variance, which might become less important for improvement in sugarcane yield. Therefore, based on PCA, it can be concluded that single cane weight, leaf area index, stalk height (SH), internodal length (IL) total biomass and internodes per stalk had positive impact on sugarcane yield in the first three components, and these characters could be useful and desirable tools for improvement in yield. The findings of this study differ from another study by Zhou *et al.*, (2015), in which PCA was applied to examine nine quantitative traits of GT sugarcane germplasm. Four principal components were derived, namely sugar component, stalk diameter and leaf factor, millable canes leaf component and stalk height component, respectively.

Table 2: Correlation matrix of the variables.

	SH	SD	SCW	NIS	IL	LAI	NTP	CY	CCS	TB
SH	1									
SD	–0.148	1								
SCW	0.006	0.824	1							
NIS	0.186	0.085	0.094	1						
IL	0.648	–0.109	0.159	–0.380	1					
LAI	–0.015	0.490	0.587	0.728	–0.17	1				
NTP	–0.027	0.169	0.254	0.138	0.008	0.42	1			
CY	0.11	0.671	0.766	0.467	0.025	0.753	0.50	1		
CCS	–0.318	0.044	0.185	–0.445	0.053	–0.225	–0.340	–0.2	1	
TB	0.261	–0.151	–0.244	–0.045	0.295	–0.304	–0.017	–0.002	–0.5	1

SH: Stalk height (cm); SD: Stalk diameter (cm); SCW: Single cane weight; NIS: No. of internodes per stalk; IL: Internodal length; LAI: Leaf area Index; NTP: No of tillers per plant; CY: Cane yield (tons ha-1); CCS: Commercial cane sugar; TB: Total biomass

Table 3: Eigen analysis of the variables.

Variable	Eigen value	Variance percentage	Cumulative percentage
SH	3.5596	35.6	35.6
SD	2.1452	21.5	57.1
SCW	1.7288	17.3	74.4
NIS	1.0131	10.1	84.5
IL	0.8321	8.3	92.8
LAI	0.3308	3.3	96.1
NTP	0.1944	1.9	98
CY	0.1157	1.2	99.2
CCS	0.0684	0.7	99.9
TB	0.0120	0.1	100

SH: Stalk height (cm); SD: Stalk diameter (cm); SCW: Single cane weight; NIS: No. of internodes per stalk; IL: Internodal length; LAI: Leaf area Index; NTP: No of tillers per plant; CY: Cane yield (tons ha-1); CCS: commercial cane sugar; TB: Total biomass

Table 4: Structure of first four principal components.

Variable	PC1	PC2	PC3	PC4
SH	0.001	**0.500**	–0.287	–0.482
SD	0.391	–0.195	–0.240	0.208
SCW	**0.418**	–0.161	–0.391	0.014
NIS	0.304	0.197	**0.446**	–0.439
IL	-0.077	**0.317**	–0.613	–0.130
LAI	**0.477**	0.004	0.148	–0.235
NTP	0.278	0.147	0.052	0.449
CY	**0.491**	0.105	–0.108	0.089
CCS	–0.126	–0.527	–0.307	–0.232
TB	–0.110	**0.486**	–0.039	0.443
Eigen value	3.560	2.145	1.729	1.013
Variance percentage	35.6	21.5	17.3	10.1
Cumulative percentage	35.6	57.1	74.4	84.5

SH: Stalk height (cm); SD: Stalk diameter (cm); SCW: Single cane weight; NIS: No. of internodes per stalk; IL: Internodal length; LAI: Leaf area Index; NTP: No of tillers per plant; CY: Cane yield (tons ha-1); CCS: Commercial cane sugar; TB: Total biomass.

The Scree plot: Scree plot helps us in deciding the number of components to retain (Lukibisi and Lanyasunya, 2010). The eigen values, associated with each component, are plotted and then searched for a breakdown between the components with large eigen values and those with smaller eigen values. The components that come before the break are considered to be important and are retained, while the components that come after the break are taken to be unimportant and are therefore not retained. In Fig. 1, it can be seen easily the break after four components and it is approved that these four components are meaningful and important.

The Score plot: The score plot is used to interpret relation among observations (Fig. 2). This graph is plotted with scores corresponding to PC1 on the X-axis and against the scores belonging to PC2 on Y-axis. Majority of the varieties lies on the left side of the plot. It is also clear that PC1 and PC2 have been effective in separating the varieties. The scores on principal component one represent greater variation and the scores with principal component two show the second largest variation. These scores are considered important in data.

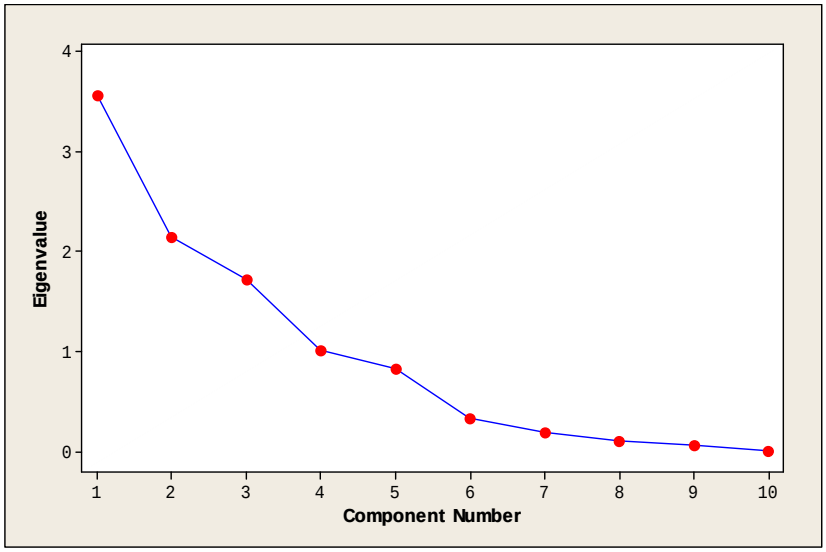

Fig. 1. Scree plot of components versus eigen values.

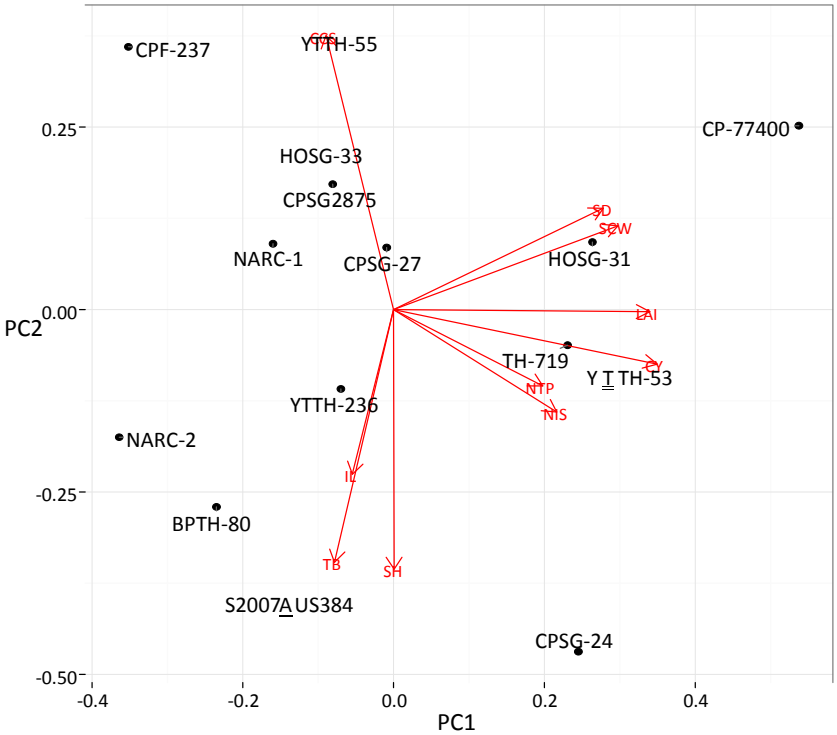

Fig. 2: Score plot for first and second component.

CONCLUSION

In the present study, an attempt has been made to prove the significance and utility of principal component analysis in analyzing agricultural data. Using PCA, ten variables of sugarcane including stalk and yield characters were reduced to transform a new set of uncorrelated variables also called principal components. Four components namely PC1, PC2, PC3 and PC4 were retained and they explained 84% of total variance in data**.** Therefore, based on PCA, it can be concluded that single cane weight, leaf area index, stalk height (SH), internodal length (IL) total biomass and internodes per stalk had positive direct effect on sugarcane yield in the first three components, and that these characters may be useful and desirable tools for improvement in sugarcane yield.

PCA can be further used in analyzing multivariate data generated from different experiments that are being carried out in all institutes of NARC. It might be an effective tool for finding the varietal indices that is the most suitable variety that could be selected by plant breeders. This technique can be disseminated among social and biological scientists in terms of conducting trainings and seminar.

REFERENCES

Ahmadzadeh, A.R., E. Majidie, B. Alizade, A.H. Omidi, 2010. Study the yield, yield components and morphological traits in the spring safflower using multivariate statistical approaches. J. New Agric. Sci., 6: 1-8.

Ayana, A. and E. Bekele, 1999. Multivariate analysis of morphological variation in sorghum germplasm from Ethiopia and Eritrea. Genet. Resour. Crop Evol., 46: 273-284.

Bahmankar, M., D.A. Nabati, M. Dehdari, 2014. Relationships among morpho-phenological traits using principal components analysis in safflower. J. Biodivers. Environ. Sci., 4(2): 89-93.

Bharadwaj, C., C.T. Satyavathi and D. Subramanyam, 2001. Evaluation of different classifactory analysis methods in some rice (*Oryza sativa*) collections. Indian J. Agric. Sci., 71(2): 123-125.

Cross, R.J., 1992. A proposed revision of the IBPGR barley descriptor list. Theor. Appl. Genet., 84: 501-507.

GoP., 2014. Economic Survey of Pakistan, Ministry of Finance, Islamabad.

Hailu, F., A. Merker, H. Singh, G. Belay, E. Johansson, 2006. Multivariate analysis of diversity of tetraploid wheat germplasm from Ethiopia. Genet. Resour. Crop Evol., 54: 83-97.

Jolliffe, I., 2002. Principal component analysis (Second ed.): Springer.

Khattree, R. and D.N. Naik2000. Multivariate data reduction and discrimination with SAS software. SAS Institute.

Lukibisi, F.B. and T. Lanyasunya, 2010. Using principal component analysis to analyze mineral composition data. Biennial Kenya Agricultural Research Institute, Scientific Conference on Socio Economics and Biometrics.

Micó, C., L. Recatala, M. Peris and J. Sanchez, 2006. Assessing heavy metal sources in agricultural soils of a European Mediterranean area by multivariate analysis. Chemosphere, 65(5): 863-872.

Minitab version 16, 2007. Statistical Data Analysis Software.

Mozzafari, K. and A.A. Asad, 2006. Relationships among traits using correlation, principal components and path analysis in safflower mutants sown in irrigated and drought stress condition. Asian J. Plant Sci., 5: 977-983.

Nikolic, D., V. Rakonjac, D. Milatovic, M. Fotiric, 2010. Multivariate analysis of vineyard peach germplasm collection. Euphytica, 171: 227-234.

Panagiotakos, D.B., C. Pitsavos, Y. Skoumas and C. Stefanadis, 2007. The association between food patterns and the metabolic syndrome using principal components analysis: The ATTICA Study. J. Am. Diet. Assoc., 107(6): 979-987.

Raychaudhuri, S., M. Stuart and R.B. Altman, 2000. Principal components analysis to summarize microarray experiments: Application to sporulation time series. Pacific Symposium on Biocomputing.

Rotaru, A.S., I. Pop, A. Vatca and A. Cioban, 2013. Usefulness of principal components analysis in agriculture. Bull. Univ. Agric. Sci. Vet. Med. Cluj-Napoca. Hortic, 69(2): 504-509.

Shlens, J., 2014. A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100.

Singh, H.K. and B. Singh, 2013. A note on multivariate technique in agricultural experiments. Int. J. Curr. Sci., 5: 50-56.

Smith, L.I., 2002. A tutorial on principal components analysis. http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf.

Upadhyaya, H.D, L.J. Reddy, S.L. Dwivedi, S. Singh, 2009. Phenotypic diversity in cold-tolerant peanut germplasm. Euphytica, 165: 279-291.

You, Q., L. Xu, Y. Zheng and Y. Que, 2013. Genetic diversity analysis of sugarcane parents in Chinese breeding programmes using gSSR markers. Sci. World J., 2013: 1-11 (Article ID 613062).

Zhou, H., R. Yang and Y.R. Li, 2015. Principal component and cluster analyses for quantitative traits in GT sugarcane germplasm (*Saccharum* Spp. Hybrids). Int. J. Agric. Innov. Res., 3(6): 1686-1690.