Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 70(1) 2016 33-39

On the Use of Principal Component Analysis in Sugarcane Clone Selection

J. Ong'ala, D. Mwanga and F. Nuani
Kenya Agricultural and Livestock Research Organization, Sugar Research Institute, Nairobi, Kenya

Received 21 August 2015; Revised 29 March 2016; Accepted 30 March 2016

SUMMARY

In the process of phenotypic evaluation of sugarcane, many traits are simultaneously evaluated. These traits are often highly interrelated; evaluation of all these traits is costly and may not enhance selection response. In this study, we used the Principal Component Analysis (PCA) to identify representative traits for phenotypic characterization of sugarcane, and thereby to select superior clones in the breeding process. The results indicate that when PCA is used, only 10 out of 17 traits are significant in identifying the superior clones and their contribution to the selected traits is quantified.

Keywords: MANOVA, Multivariate analysis, Principal component analysis, Trait combination, Variable reduction.

1. INTRODUCTION

Sugarcane contributes to 70% of the raw material of sugar produced world-wide (Butterfield *et al.* 2001). It is also used as a bio-energy crop due to its phenomenal dry matter production (Chohan *et al*. 2014). The demand for white sugar and bio-energy is increasing day by day due to the consistent increase in population, but its present production is not enough to meet the increasing demand because of low production (Cheema and Mahmood 2005). Some of the sugarcane yield productions limiting factors amongst others are: lack of high cane and sugar yielding (Chohan *et al*. 2007) or their low rate of adoption of the available varieties.

New sugar cane varieties are produced by sexual means and propagated vegetative. Each year a new population of original seedlings consisting of thousands of new varieties is produced through fuzz (true seed). These are screened clonally through several selection stages, their numbers being reduced at each stage and the selected ones tested in larger plots in which their performance can be evaluated more reliably. The time taken to release sugarcane variety ranges from eight to twenty years (Tai, *et al*. 1992) making the monetary and time resources involved in generating a new variety quite immense.

During phenotypic evaluation of sugarcane clone, many traits are simultaneously evaluated, which are often genetically linked. It is costly to evaluate all the traits which probably may be interrelated and does not ensure optimal selection gains. The appropriate methods that provide accurate evaluations and estimation of genetic diversity depends on genetic variation, sampling methods, the magnitude of data sets, and the statistical tools applied in the data analysis (Mohammadi and Prasanna 2003). The two important characters for obtaining high sugar yield are cane yield and sucrose content (Terzi *et*

Corresponding author: J. Ong'ala
E-mail address: ongalajacob@gmail.com

al. 2009) and therefore cane yield and sucrose content and their interaction are important parameters for developing superior genotypes (Zhu *et al.* 2000, Chohan *et al.* 2007, Alvarez *et al*. 2009). A study by Olaoye (1995) observed that four characters viz., field emergence, stalks/stool, stalk length and stalk diameter could account for 31 to 53% variation in cane yield and sucrose content. Another study by Khan *et al*. (2012) revealed that higher number of tillers, good weight, endowed with better available sugar in the cane (pol)%, commercial cane sugar (CCS)% and purity% are the important characters which should be considered in selection of higher sugar yield in sugarcane genotypes.

The principal component analysis (PCA) is one powerful statistical method widely applied to classify phenotypic traits in crop germplasm into groups based on similarities (Rukundo *et al.* 2015). The purpose of principal component analysis is to find the best low-dimensional representation of the variation in a multivariate data set. We can carry out a principal component analysis to investigate whether we can capture most of the variation between samples using a smaller number of new variables (principal components), where each of these new variables is a linear combination of all or some of the traits.

PCA reduces the original variables into a new set of uncorrelated variables known as principal components (PCs). These PCs clarify the connections between traits and divide the total variance of the original traits into a small number of uncorrelated new variables (Wiley and Lieberman 2011). The PCA allows visual differentiation among entries and identify possible associations (Mohammadi and Prasanna 2003) by providing a two dimensional scatter plot consisting of individual entries. The geometrical distances among individuals in this plot reveal the genetic distances among them. Amalgamation of individuals in a similar quadrant of a plot may indicate a group of genetically related individuals (Warburton *et al*. 2002, 1832-1840). The objective of this study therefore was to use principal component analysis to identify principal traits for efficient phenotypic characterization of sugarcane in identifying superior clones for release.

2. MATERIALS AND METHODS

2.1 Planting Material

Sugar Research Institute (SRI) has maintained, since the 1980s, more than 100 parents that were selected from the inter-specific programmes across the world. This collection was used in the hybridization programme for broadening the genetic base of varieties. The parents that were used in this experiment were KEN82-401, CO462, CO513, CO1148, CO285, CO746, CO527, CO945, CO421, CO617, N14,N52-219, EAK71-402, EAK70-97, EAK75272, B41211. About 30,000 seedlings were obtained for stage 1 of this trial. After hybridization, sugarcane seedlings (each seedling is apotential variety/clone) were raised from true seed in stage 1 and in the subsequent stages, sets were used for planting. The evaluation of the clones stage by stage was perfomed against the standard commercial varieties (N14, CO421), and selection done using the conventional methods from stage 1 to stage 4. In stage 3, only following clones viz: 01-374, 10-496, 105-11, 105-13, 25-1097, 25-1104, 25-1136, 58-704 performed better than the standards checks and were advanced to stage 4. The seven clones together with the 2 standard checks were used as the planting materials in stage 4. Therefore this paper is based on the data that was generated in stage four of the selection cycle.

2.2 Site

In the first stage, seedlings at Mtwapa breeding centre ($3^{o}56$'S, $39^{o}44$'E, 15m above sea level). In stage 2 and stage 3 of the experiment was performed at three ecological zones Mtwapa, Kibos (0° 4' 0S34° 49' 0E, 1,135 m above sea level) and Kikoneni (4° 27' 0S, 39° 17' 60E, 77 m above the sea level) while the trial at stage 4 was conducted in Kikoneni, Kibos, Kwale International Sugar Company Limited (KISCOL) (4° 31' 60S, 39° 22' 60E, 8m above sea level), Mumias (0° 20' 11N, 34° 29' 21E, 1268) and South Nyanza (1° 4' 0S, 34° 28' 0E, 1322m above the sea level). The sites mentioned above are the major sugarcane growing areas. In this paper however we have used the data from Kibos site to demonstrate the use of PCA in sugarcane selection process.

2.3 Design

The experimental designs that were used in stage 4 was the Randomized Complete Block Design (RCBD). Once the field for laying the experiments were identified, it was sub-divided into three blocks (homogeneous within blocks). The blocks were then subdivided into nine plots (each measuring 10.5m × 8m). Each plot represented the experimental unit.

2.4 Data Collection

Observations were recorded for the 17 important agronomic characters viz;, germination at 30 days after planting (G30D), 45 days after planting (G45D), tiller count at 3 months after planting (T3M), 5 months after planting (T5M), 7 months after planting (T7M), Brix reading at 12 months after planting (Brix12M), 13 months after planting (Brix13M), 14 months after planting (Brix14M), 15 months after planting (Brix15M), 16 months after planting (Brix16M), cane plot weight (CANE_PLOT_WGT), population (POP), girths, height, number of internodes (INT), Pol percent cane (POL), fibre, purity and tons of cane per hector (TCH) which was the computed from the cane plot weight.

2.5 Data Analysis

Data analysis was done using R software on the experimental data mentioned above. All collected quantitative data were subjected to the standard analysis of variance and Multivariate Analysis of Variance (MANOVA) to measure the group effect of the parameters clone differences. A principal component analysis (PCA) was also carried out to identify and classify genotypes and to identify principal traits to be used in phenotyping sugarcane variety.

3. RESULTS AND DISCUSSION

3.1 Correlations Analysis

A correlation analysis of all the traits was performed and the coefficients and P-Values of the significant relationships are shown in Table 1. The result indicates very strong positive correlations between and among the germination and tiller counts of different periods after planting. A strong negative correlation is seen between Pol and fibre (Coeff = -0.70086, P-Value = 4.76E-07). These significant correlations between traits give early indication that some traits might just be represented by other trait when deciding on superior clones. The rest of the pairs of traits were not significant hence not included in the table.

Table 1. Correlation analysis of the sugarcane traits

Trait i	Trait j	Coeff	P-Value	i	j	Coeff	P-Value
G30D	G45D	0.859446	1.25E-12	Brix14M	GIRTH	-0.42553	6.19E-03
G30D	T3M	0.795109	8.90E-10	Brix15M	GIRTH	-0.43939	4.56E-03
G45D	T3M	0.826617	5.01E-11	Brix16M	GIRTH	0.399619	1.06E-02
G30D	T5M	0.696076	6.13E-07	POP	GIRTH	-0.48308	1.59E-03
G45D	T5M	0.720958	1.55E-07	Brix12M	INT	0.333599	3.54E-02
T3M	T5M	0.791208	1.23E-09	GIRTH	INT	0.454239	3.24E-03
G30D	T7M	0.801529	5.16E-10	HGT	INT	0.371721	1.82E-02
G45D	T7M	0.845837	6.41E-12	POL	FIBRE	-0.70086	4.76E-07
T3M	T7M	0.832388	2.78E-11	Brix13M	PURITY	-0.33379	3.53E-02
T5M	T7M	0.914092	2.22E-16	POL	PURITY	0.55063	2.32E-04
Brix12M	Brix14M	-0.32468	4.09E-02	G45D	TCH	0.382192	1.49E-02
Brix13M	Brix15M	0.313261	4.90E-02	T3M	TCH	0.411727	8.30E-03
Brix14M	Brix15M	0.368039	1.95E-02	GIRTH	TCH	0.338846	3.25E-02
Brix13M	Brix16M	-0.33101	3.69E-02	HGT	TCH	0.640667	8.47E-06
Brix15M	Brix16M	-0.38332	1.46E-02	INT	TCH	0.394552	1.18E-02
Brix16M	POP	-0.41136	8.36E-03				

Table 2. Multivariate analysis of variance

Model	Traits Included	MANOVA Statistics	
		Pillai Test	P-Value
1.	G30D, G45D, T3M, T5M, T7M, Brix12M, Brix13M, Brix14M, Brix15M, Brix16M, POP, GIRTH, HGT, INT, POL, FIBRE, PURITY, TCH	5.4578	5.379e-05 ***
2.	G30D, Brix12M, Brix13M, Brix16M, POP, GIRTH, HGT, INT, POL, FIBRE, PURITY,TCH	4.0965	3.101e-05 ***
3.	G30D, Brix12M, Brix16M, POP, GIRTH, HGT,INT, POL, FIBRE, PURITY, TCH	3.9849	5.390e-06 ***
4.	G30D, Brix16M, POP, GIRTH, HGT, INT, POL, FIBRE, PURITY, TCH	3.8947	5.515e-07 ***
5.	G30D, Brix16M, POP, GIRTH, HGT, INT, POL, FIBRE, TCH	3.6714	3.231e-07 ***

3.2 Multivariate Analysis of Variance (MANOVA)

Having observed the correlation between the traits, a MANOVA (Carey 1998) was performed to show the group effect of all the traits on the tested clones. Whilst excluding the trait that had significant correlation with the others already existing in the model, further MANOVA was performed and the models evaluated. The results of MANOVA are shown in Table 2. In model 1, all the traits that were measured had a significant group effect on the clones. In model 2, some traits (G45D, T3M, T5M, T7M, Brix12M, Brix14M, Brix15M) which were found to have significant correlation with the traits used in the model were omitted. The effect of the traits in model 2 had stronger significance than that in model 1 (P-value for model 2 = 3.101e-05 < P-value for model 1 = 5.379e-05) suggesting that the differences in clones is more evident when fewer traits are used. Model 5 had the lowest number of traits and it gave the strongest evidence of difference in clone. All the traits in model were not significantly correlated. This suggests that correlation of traits contributes to their redundancy in evaluating clone.

3.3 Principal Component Analysis (PCA)

In the evaluation of diversity among the sugarcane clones using 19 traits, it is observed that the first three components (PCA1, PCA2 and PCA3) explained upto 80.8% of the total variation among traits (See Table 3). On interpretation of the rotation, the first principal component (PCA1) is more (farthest from zero) related negatively with G45D, T3M, T5M and T7M and positively correlated with Brix16M.

This indicates that G45D, T3M, T5M and T7M vary together but inversely with Brix16M. It can be noted that the same traits (G45D, T3M, T5M and T7M) with negative loadings with PCA1 ware not important including them in the combined effect on clone (MANOVA model 5) in Table 2. The same interpretation follows with PCA2 and PCA3. PCA2 was important in selecting high yielding clones in terms of TCH while PCA 1 was important in selecting high sucrose clones (or early maturing) in terms of Brix16M as shown in Table 4.

This classification is based on G45D, T3M, T5M, T7M, Brix16M, and TCH. The clones in quadrant one of Fig. 1 (25_1097, CO421, 105_13, 58_704) can be classified together in terms of their similarities on Brix16M which had a positive coefficient for PCA1 and Brix14M which had a negative coefficient for PCA2. The clone named 10_496 is classified on its own (appearing itself in quadrant 2 of Fig. 1) based on high on Brix16M and TCH with PCA1 and PCA2 respectively. The genotypes appearing in quadrant (1) and (2) are characterised with high TCH and girth while those appearing in quadrant (2) and (4) are characterized with high brix at the age of 16 months. A more illustrated plot for clones when data points for different replication are used is shown in Fig. 2.

3.4 Classification of Clones using PCA1 and PCA3

This classification is based on G45D, T3M, T5M, T7M, Brix16M, Height, INT and PURITY. Their results are shown in Fig. 3. Using high purity and high BRIX16M, PCA grouped N14, 25_1104 and 10_496 in one cluster.

The genotype 10_496 was the best in the brix at age of 16 months rate and yield in TCH (Fig. 1 and Table 5). The traits shown in Table 5 were having positive loadings to the principal components that accounted for 80.8% of the total variation within the genotype hence most important in the evaluation process.

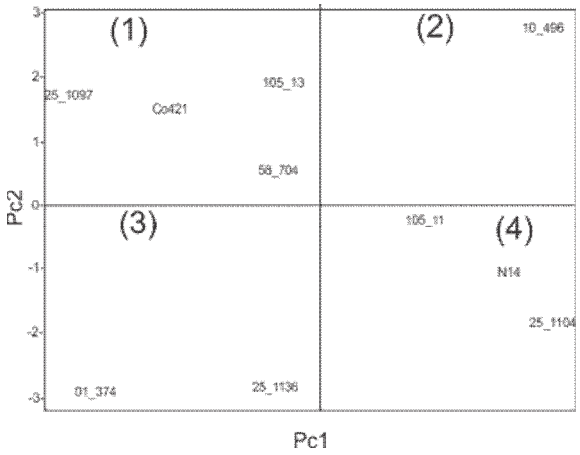

Fig. 1. PCA1 by PCA2 bi-plot showing average performance of clones

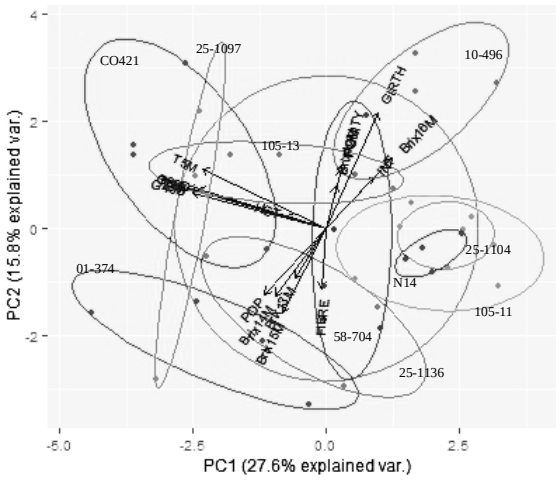

Fig. 2. Detailed PCA1 by PCA2 bi-plot showing average performance of clones

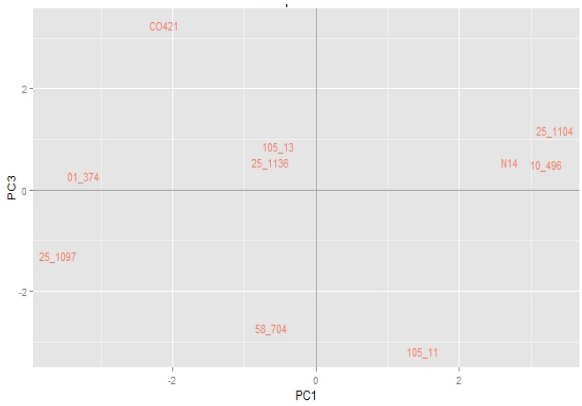

Fig. 3. PCA1 by PCA3 plot showing average performance of clones

Table 3. Importance of Components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	2.5833	2.0511	1.9141	1.04427	0.99638	0.81733	0.57931	0.50468	0.33789	2.23E-16
Proportion of Variance	0.3708	0.2337	0.2035	0.06058	0.05515	0.03711	0.01864	0.01415	0.00634	0.00E+00
Cumulative Proportion	0.3708	0.6045	0.808	0.8686	0.92375	0.96086	0.97951	0.99366	1	1.00E+00

Table 4. Rotations

Trait	PC1	PC2	PC3	Trait	PC1	PC2	PC3
G30D	-0.2880	0.2246	0.2138	Brix16M	0.3094	0.1852	0.1514
G45D	-0.3006	0.2458	0.1603	POP	-0.2420	-0.1740	0.0048
T3M	-0.3287	0.1746	0.1868	GIRTH	0.1833	0.3637	0.0192
T5M	-0.3008	0.1579	0.2645	HGT	-0.2037	0.0191	-0.4018
T7M	-0.3218	0.1795	0.2130	INT	0.1055	0.2008	-0.3925
Brix12M	-0.0281	0.3677	-0.2687	POL	0.2427	-0.1296	0.2130
Brix13M	-0.2497	-0.2740	-0.1079	FIBRE	-0.1641	-0.1134	-0.2683
Brix14M	-0.2032	-0.3295	0.0921	PURITY	0.1480	-0.0090	0.4149
Brix15M	-0.2250	-0.2742	-0.1156	TCH	-0.1461	0.3745	-0.2107

Table 5. Means of traits highly correlated with pc1, pc2, and pc3 for 10 evaluated sugarcane genotypes

	Brix16M	POP	GIRTH	PURITY	TCH
01_374	20.71	592.75	1.93	89.26	137.03
10_496	21.20	500.00	2.85	89.71	158.23
105_11	20.96	505.50	2.56	88.62	159.35
105_13	21.13	584.50	2.41	89.47	166.33
25_1097	20.68	560.50	2.54	88.02	188.65
25_1104	21.18	514.25	2.59	89.40	123.63
25_1136	20.73	578.50	2.20	89.28	133.96
58_704	20.95	560.25	2.42	88.79	160.78
CO421	21.03	522.75	2.63	90.08	152.19
N14	21.14	554.25	2.34	89.86	147.09

4. CONCLUSION

In conclusion, the genotypic variance component made the greatest contribution to the sources of variation among the test genotypes for G45D, T5M, T7M, Brix16M, Height, INT and, population, girth, purity and TCH. Therefore, potential clones could be identified for selection. Among the 19 phenotypic traits used in the current genetic diversity study, the PCA identified only 10 phenotypic traits, contributing to 80.8% of variations. There are some similarities between the results multivariate analysis of variance and PCA in terms of the traits that were not important for evaluating genotypes. PCA has been applied in this work as a statistical approach to identify major variance components, their contributions and correlated traits. This method assists in reducing the number of traits in the data collection in a breeding and selection process. In this method, fewer variables explaining variations among individuals can be screened among various traits. Therefore, the PCA provides valuable information when there are several correlated traits, by reducing the costs of screening. Overall, the clones 10_496 was identified as good varietal candidate for high yielding (TCH) and quality (Brix at 16 Months).

ACKNOWLEDGEMENTS

I wish to acknowledge SRI breeders; Dr. Jamoza, Ong'injo, and Ochia for their contribution in the field experiments and set up, Thanks also to my field assistants; Kiratu, Ngutu and Ocham. I thanks millers; KISCOL, MSC and SONY for providing sitesduring the experiment and finally my sincirere thanks to KALRO-SRI for providing resources for the study.

REFERENCES

Alvarez, J., Deren, C.W., and Glaz, B. (2009). Sugarcane Selection for Sucrose and Tonnage Using Economic Criteria. Florida Cooperative Extension Service, UF/IFAS, University of Florida, Gainesville, FL. Published May 2004, Reviewed August 2009.

Butterfield, M., D'Hont, A. and Berding, N. (2001). The sugarcane genome: a synthesis of current understanding, and lessons for breeding and biotechnology. *Proc. Soc. Afr. Sugarcane Technol. Ass.*, **75**, 1-5.

Carey, Gregory (1998). *Multivariate Analysis of Variance (MANOVA)*: I. Theory, Colorado.

Cheema, I.A. and Mahmood, T. (2005). Efficacy of different herbicides to control weeds in spring planted sugarcane. *Pak. Sugar J.*, **20(5),** 5-8.

Chohan, M., Panhwar, R.N., Memon, M.A., Unar, G.S. and Mari, A.H. (2007). Performance of new sugarcane varieties for cane yield and quality under different agro-climatic conditions of Sindh. *Pak. J. Sci. Res.*, **59,** 28-33.

Khan, I.A., Bibi, S., Yasmin, S., Khatri, A., Seema, N. and Abro, A.S. (2012). Correlletion studies of the agronomic traits for higher sur yield in sugarcane. *Pak. J. Botany*, **44 (3),** 969-971.

M. Chohan, U.A., Talpur, S., Junejo, G.S., Unar, R.N. and B., P.A. (2014). Selection and evaluation of the diverse sugarcane genotypes in 4th stage. *The J. Anim. Plant Scis.*, **24(1),** 197-203.

Mohammadi, S. and Prasanna, B. (2003). Analysis of genetic diversity in crop plants salient statistical tools and considerations. *Crop. Sci.*, **43(4),** 1235-1248.

Olaoye, G. (1995). Evaluation of local sugarcane accession II determinants of cane yield and sucrose content. *Nigeria J. Genet.*, **10,** 23-30.

Rukundo, P., Hussein, S., Mark, L. and Daphrose, G. (2015). Application of principal component analysis to yield and yield related traits to identify sweet potato breeding parents. *Tropical Agric. (Trinidad),* **92(1).**

Tai, P.Y., Shine, J. and Miller, J. (1992). Cross evaluation using a small progeny test. *Amer. Soc. Sugarcane Tech.*, **12,** 110-111.

Terzi, F.S., Rocha, F.R., Vencio, R.Z., Felix, J.M., Bran, D.S., Waclawovsky, A.J., *et al.* (2009). Sugarcane gene associated with sucrose content. Retrieved from BMC Genomics: http://www.biomedcentral.com/1471-2164/10/120

Wiley, E. and Lieberman, B. (2011). *Phylogenetics: Theory and Practice of Phylogenetic Systematics.* John Wiley and Sons, New York.

Zhu, Y.J., Albert, H.H. and Moore, P.H. (2000). Differential expression of soluble acid invertase genes in the shoots of high-sucrose and low-sucrose species of Saccharum and their hybrids. *Austr. J. Plant Physiol.*, **27,** 193-199.