STEPWISE AND CANONICAL DISCRIMINANT ANALYSIS OF MINERAL COMPOSITION DATA

Lukibisi, F.B, Muia, J.M.K, and Lanyasunya, T.

National Animal Husbandry Research Centre,P.O. Box 25, Naivasha

ABSTRACT

Discriminant analysis is one of the classical classification techniques used to discriminate a single categorical variable using multiple attributes. Discriminant analysis also assigns observations to one of the pre-defined groups based on the knowledge of the multi- attributes. When the distribution within each group is multivariate normal, a parametric method can be used to develop a discriminant function using a generalized squared distance measure. The classification criterion is derived based on either the individual within-group covariance matrices or the pooled covariance matrix that also takes into account the prior probabilities of the classes. The performance of a discriminant criterion could be evaluated by estimating probabilities of mis-classification of new observations in the validation data. A user-friendly SAS application utilized to perform discriminant analysis is presented here. Mineral composition data containing multi-attributes is used to demonstrate the features of discriminant analysis in discriminating the four sites.

Key words: Canonical discriminant analysis (CDA); Discriminant analysis (DA); Mahalonobis distance; Stepwise discriminant analysis.

INTRODUCTION

There are several methods to measure variation. With univariate analyses, each variable is analyzed separately allowing for substantial overlapping of results. Univariate statistical techniques such as analysis of variance do not explain how regions differ when all measured variables are considered jointly. In canonical discriminant analysis, a multivariate statistical technique, all variables are considered simultaneously in the differentiation of populations. This approach results in a more powerful comparison of populations than can be achieved with univariate analysis, provided the variables are correlated. Canonical descriminant analysis can separate among population effects from within population effects by maximizing discrimination among populations when tested against the variation within populations (Riggs, 1973). After determination of the among population variability, the Mahalanobis distance (D*) statistic can be used as an indicator of the difference between populations (Boles and Swan, 2002; McDonagh et al., 2001). The information obtained from CDA can then additionally be used to group the regions-populations into smaller subgroups that are more similar to each other (Khattree and Naik, 2000; Lukibisi, et al 2008). Multivariate procedures based on phenotypic and agronomic characteristics have been used in the assessment of genetic diversity in soybean (Bains and Sood, 1984), perennial ryegrass (Humphreys, 1991), weed (Kenkel, et al., 2002) , and tall fescue (Vaylay and van Santen, 2002).

In this paper we present the more commonly used multivariate scaling methods, including descriptive and predictive models. Our objective is to instill in the researchers an intuitive understanding of these methods and their applications using numerical examples from mineral composition data.

MATERIALS AND METHODS

The study was conducted at the National Animal Husbandry Research Centre (NAHRC) in Naivasha, Kenya over a period of one year. The browse material (buds, pods, leaves and twigs of less than 5mm) to be evaluated was collected from four different sites (Samburu, Njoro, Nyandarua and Naivasha). The herbage was sampled for both wet and dry seasons. Within, the four sites, 6 suitable sites were identified and within each site, 6 sampling plots (100 x 100 metres in size) were demarcated. Bulk of the collected material from each site stratified per sampling plot and sections of the plant harvested, were air-dried and further divided into two portions (1:2) for degradability and chemical analysis respectively. Small representative samples of fresh materials were also transported in cool boxes to the research laboratory and

immediately oven dried at 105°C for 24 hours to determine the dry matter (DM). Chemical analyses: The material used for chemical analyses were further oven dried and milled pass 1 mm sieve and packed (air-tight) in strong polythene bags. Ash was

The method extracts n-1 discriminant functions, n being the number of groups to discriminate among, which are linear combinations of the original quantitative variables selected. These functions may be used to calculate a set of discriminant scores that

TABLE I- STEPWISE SELECTION SUMMARY FOR DATA SET ONE

Step	Number in	Character*	Partial R-square	F-value	Pr>F	Wilks' Lambda	Pr <Lambda	Average squared canonical correlation	Pr>ASCC
1	1	Cu	0.7205	24.06	0.0001	0.279510	<0.0001	0.240163	<0.0001
2	2	Fe	0.3813	5.55	0.0042	0.172946	<0.0001	0.333019	<0.0001
3	3	Mg	0.3444	4.55	0.0108	0.113377	<0.0001	0.416195	<0.0001
4	4	Ca	0.4866	7.90	0.0007	0.058212	<0.0001	0.482177	<0.0001
5	5	P	0.3372	4.07	0.0180	0.038585	<0.0001	0.558745	<0.0001
6	6	K	0.3062	3.38	0.0354	0.026770	<0.0001	0.606168	<0.0001
7	7	Zn	0.2583	2.55	0.0815	0.019856	<0.0001	0.631756	<0.0001

determined (AOAC, 1990, ID 942.05) and micro-Kjeldahl N in feed (AOAC, ID 954.01). Crude protein was calculated as Kjeldahl Nx6.25. Neutral-detergent fibre (NDF), acid-detergent fibre (ADF) and acid-detergent lignin (ADL) were determined by the procedures of Van Soest and Robertson (1985). Ether extract (EE) was determined by extracting the sample with petroleum ether using a Gerhart Soxtherm 2000 Automated (AOAC, 1990, ID 920.39). Representative samples taken from the bulk feed material (dried at 40oC and grounded to pass 1 mm screen) were also assayed for total extractable phenolics (TEP) by the Folin assay method (Folin-Denis method) as described by Waterman and Mole (1994) and Pallm and Rowland (1997). Linear regression was used to contruct a standard curve relating absorbance (760 nm) of the Folin assay reaction mixture to the known concentration of tannic acid (µg) (standard). Similarly, representative dry samples (approx. 5 µg) from forage material was also assayed for both essential and non-essential amino acids (AA) concentration according to AOAC (1990, ID 994.12 Llames and Fontaine 1994). Sodium (Na) and Potassium (K) were determined by flame photometry. Phosphorus (P) was determined by spectrophotometry. Callcium (Ca), Magnesium (Mg), Manganese (Mn), Copper (Cu) and Zinc (Zn) were determined by AAS (Atomic absorption spectrophotometer) (AOAC, 1990).

Statistical analysis

Discriminant analysis is a well-known multivariate statistical classification technique used to determine which variables discriminate between two or more groups, given several quantitative (independent) variables and a categorical (dependent) variable. are employed to predict the status of a new observation. In this study a forward stepwise procedure, guided by the respective ''F to enter'', has been applied to reduce the original number of variables. The F value indicates the statistical significance of a variable in the discrimination among the four groups (Morrison, 1976). The model parameters are Wilks' Lambda, an index of the discriminating power ranging between 0 and 1 (the lower the value the higher its discriminating power): eigenvalues, a measure of the variance in the dependent variable for each function: canonical correlations, a measure of the association between the groups formed by the dependent and the given discriminate function (the larger this value, the higher in the correlation between the discriminant function and the group). The first discriminant function (DF), orthogonal to the first, maximizes the residual differences between values of this variable. And so on. The DF1 will be the most powerful differentiating dimension, but later functions may also represent additional significant dimensions of differentiation. Because of the different size of the groups under study, the predictions were accordingly adjusted using a priori probabilities classification. The predictive validity of the model has been assed by a leave-one-out cross validation method.

RESULTS AND DISCUSSION

Stepwise discriminant analysis found most of the variables to be significant ($P<0.001$) with respect to their partial r^2 (table 1) for data set one. The most important variable for discriminating the sites was Cu with the partial r^2 of 0.72, followed by Fe, Mg, Ca, P, K, Zn and for data set two (table 2), the most important variable was Cu with the partial r2 of 0.88

TABLE II- STEPWISE SELECTION SUMMARY FOR THE DATA SET TWO

1	1	Cu	0.8844	12.26	0.0001	0.115557	<0.0001	0.294815	<0.0001
2	2	Ca	0.3088	6.40	0.0011	0.079866	<0.0001	0.396088	<0.0001
3	3	P	0.3048	6.14	0.0015	0.055525	<0.0001	0.420472	<0.0001
4	4	N	0.2346	4.19	0.0113	0.042500	<0.0001	0.436544	<0.0001
5	5	Mn	0.1551	2.45	0.0778	0.035910	<0.0001	0.474969	<0.0001
6	6	Mg	0.1626	2.52	0.0717	0.030072	<0.0001	0.505796	<0.0001

followed by Ca, P, N, Mn and Mg. Thus, the variables in stepwise discriminant analysis were selected in that order respectively in the two data sets.

The correct classification percentages in each region (1, 2, or 3) were calculated as well from Table 3 and T4 for data sets one and two respectively.

TABLE III- LINEAR DISCRIMINANT FUNCTION FOR SITE IN DATA SET ONE

	SITES			
Variable	1	2	3	4
Constant	-19.923	-15.339	-6.971	-9.812
P	-12.113	37.001	37.532	38.816
K	-3.760	-2 324	-0.797	-0.404
Ca	20.826	6.740	-4.229	1.724
Mg	-36.132	-14.956	12.822	0.995
Fe	0.041	0.014	-0.014	-0.002
Mn	0.002	0.024	0.015	0.007
Zn	0.185	0.096	0.079	0.050

Site1 = -19.923 – 12.113*P – 3.760*K + 20.826*Ca – 36.132*Mg + 0.041*Fe + 0.002*Mn + 0.185*Zn
Site2 = -15.339 + 37.001*P – 2.324*K + 6.740*Ca -14.956*Mg + 0.014*Fe + 0.024*Mn + 0.096*Zn
Site3 = -6.971 + 37.532*P – 0.797*K – 4.229*Ca + 12.822*Mg – 0.014*Fe + 0.015*Mn + 0.079*Zn
Site4 = -9.812 + 38.816*P – 0.404*K + 1.724*Ca + 0.995*Mg - 0.002*Fe + 0.007*Mn + 0.050*Zn

TABLE IV- LINEAR DISCRIMINANT FUNCTION FOR SITE IN DATA SET TWO

	SITES			
Variable	1	2	3	4
Constant	-46.765	-11.420	-58.567	-11.938
N	13.483	6.826	12.980	7.643
P	-23.799	3.028	-37.708	-0.003
Ca	11.731	7.054	4.988	6.778
Mg	-1.608	-4.957	9.803	-1.388
Cu	1.619	0.454	1.968	0.503
Mn	-0.025	-0.002	0.001	-0.010

Site1 = -46.765 +13.485*N – 23.799*P + 11.753*Ca – 1.608*Mg + 1.619*Cu - 0.025*Mn
Site2 = -11.420 + 6.826*N + 3..028*P + 7.054*Ca – 4.957*Mg + 0.454*Cu - 0.002*Mn
Site3 = -58.567 + 12.980*N – 37.708*P + 4.988*Ca + 9.803*Mg + 1.968*Cu + 0.001*Mn
Site4 = -11.938 + 7.643*N – 0.003*P + 6.778*Ca – 1.388*Mg + 0.503*Cu - 0.010*Mn

The above equations can be used to estimate the missing value in the respective sites for data set one.

The above equations can be used to estimate the missing value in the respective sites for data set one.

TABLE V- MATRIX STRUCTURE COEFFICIENT, PERCENTAGE OF VARIANCE, EIGNVALUES, CANONICAL CORRELATIONS AND WILKS' LAMBDA OF THE FINAL MODEL FOR DATA SET ONE

	Function		
Variable	1	2	3
P	0.1328	0.8266	-0.1402
K	-0.0896	0.5057	0.4178
Ca	0.2773	0.5918	0.3061
Mg	-0.0699	0.4767	0.2473
Fe	0.2688	0.6222	0.9575
Mn	0.0072	0.0951	-0.4555
Zn	0.0740	-0.1811	-0.0319
% of variance	75.83	15.75	8.42
Eigenvalues	5.079	1.055	0.564
Canonical Correlation	0.914	0.717	0.600
p-value	0.0001	0.0033	0.0447
Wilks' Lambda		0.0512	

To eliminate the variables that provided superfluous information at a 99% level, a F to enter=8 in the forward stepwise procedure with tolerance of 0.05 was applied. Because of the four levels of the categorical variable, four significant discriminant functions of classification were obtained (Tables 5 & 7). Moreover, seven variables, i.e. Cu, Fe, Mg, Ca, P, K and Zn were selected for data set one and six elements, i.e. Cu, Ca, P, N, Mn and Mg were also selected for data set two shown in Tables 1 & 2 respectively. Model parameters, in terms of percentage of explained variance, eigenvalues, canonical correlation values, p-values and Wilks' Lambda values of 0.0511 ($p<0.0001$) for data set one and 0.0301 ($p<0.0001$) for data set two respectively shown in Tables 5 and 7 showed good discriminant power of the model.

The CAN1 explained 75.83% of the total variance with a good correlation value (0.914) for data set one shown in Table 5 and 93.16% of the total variance with a good correlation value (0.967) for data set two shown in Table 7 respectively, therefore, it is the best discriminating between sites (see Figures 1, 2). In Tables 5 and 7, the matrix structure coefficients, showing the correlations of each variable in the model with each discriminant function were also reported respectively. The structure coefficients are global (not partial) coefficients, similar to correlation coefficients, and reflect the uncontrolled association of the discriminating variables with the categorical variable.

Table 5 lists the pooled within canonical structure coefficients for data set one. Canonical variable 1 (Can1) had the highest correlation with Ca (0.2773) followed by Fe (0.2688) and P (0.1328). Canonical variable 2 (Can2) had the greatest correlation with P (0.8266) followed by Fe (0.6222), Ca (0.5918) and K (0.5057) and Canonical variable 3 (can3) had the highest correlation with Fe (0.9575), followed by K (0.4178) and Ca (0.3061).

The variance associated with the first two canonical

TABLE VI - STANDARDIZED CANONICAL COEFFICIENTS FOR DATA SET ONE

	Function		
Variable	1	2	3
P	-0.8964	0.8843	-0.9418
K	-1.0485	0.5497	0.6648
Ca	1.6238	-0.1049	0.6012
Mg	-1.2392	-0.1718	0.1827
Fe	1.6465	-0.0024	0.1267
Mn	-0.1035	0.1165	-0.7786
Zn	0.6268	-0.6928	0.0298

variables was 91.58% of the total variance (Table 5). How the sites were grouped and separated using these three canonical variables is illustrated in Figure 1.

In Table 6, the standardized discriminant coefficients used to compare the relative importance of the independent variables were listed. The higher their absolute value, the greater is their unique contribution to the discriminant power. It was possible to asses that Fe (1.6465), Ca (1.6237), Mg (-1.2339) and K (-1.10485) were the most important discriminating variable in the Can1.

Similarly, Table 7. lists the pooled within canonical structure coefficients for data set two. Canonical variable 1 (can1) had the highest correlation with Cu (0.7272) and it is the only one with positive correlation. Canonical variable 2 (can2) had the greatest correlation with Ca (0.6541) followed by P (0.2220) and Cu (0.1086) and Canonical variable 3 (can3) had the highest correlation with Mn (0.5113), followed by Cu (0.4498) and P (0.2598). Here we also see that Mg is negatively correlated (-0.6400).

In Table 8, the standardized discriminant coefficients used to compare the relative importance of the independent variables were listed. The higher their absolute value, the greater is their unique contribution to the discriminant power. It was possible to asses that Cu (1.3277) and N (0.4925) were the most important discriminating variable in the Can1. For can2, Ca (1.2824), Mg (-0.6546), Mn (-0.6253) and N (0.5420) were most important while for can3 we have Mn (0.6439), Mg (-0.5667) and N (-0.4818) as the important variables.

The classification data for each Site were reported in Tables 9 and 10 for data sets one and two respectively. And these classifications are represented in the scattered plots of Figure 1and 2 for tall the discriminant functions

The classification data for each Site were reported in Tables 9 and 10 for data sets one and two respectively. And these classifications are represented in the scattered plots of Figure 1and 2 for tall the discriminant functions

TABLE VII- MATRIX STRUCTURE COEFFICIENT, PERCENTAGE OF VARIANCE, EIGENVALUES, CANONICAL CORRELATIONS AND WILKS' LAMBDA OF THE FINAL MODEL FOR DATA SET TWO

	Function		
Element	1	2	3
N	-0.0880	0.0297	-0.4754
P	-0.1125	0.2220	0.2598
Ca	-0.0764	0.6541	0.0127
Mg	-0.0476	0.0082	-0.6400
Cu	0.7272	0.1086	0.4498
Mn	-0.0214	-0.1243	0.5113
% of variance	93.16	6.17	0.67
Eigenvalues	14.414	0.955	0.103
Canonical correlation	0.967	0.699	0.306
p-value	0.0001	0.0004	0.3885
Wilks' Lambda	0.03007		

The variance associated with the first two canonical variables was 99.33% of the total variance (Table 7) for data set two. How the sites were grouped and separated using these three canonical variables is illustrated in Figure 2.

TABLE VIII - STANDARDIZED CANONICAL COEFFICIENTS FOR DATA SET TWO

	Function		
Element	1	2	3
N	0.4925	0.5420	-0.4818
P	-0.8934	0.2415	0.3370
Ca	0.0242	1.2824	0.0404
Mg	0.2297	-0.6546	-0.5667
Cu	1.3277	0.1765	-0.0201
Mn	-0.0150	-0.6253	0.6439

TABLE IX - CLASSIFICATION MATRIX FOR DATA SET ONE

	Predicted site membership and percentage in brackets set one				
Site	1	2	3	4	Total
1	8(100)	0(0.00)	0(0.00)	0(0.00)	8(100.00)
2	0(0.00)	8(100.00)	0(0.00)	0(0.00)	8(100.00)
3	0(0.00)	0(0.00)	8(100.00)	0(0.00)	8(100.00)
4	0(0.00)	0(0.00)	0(0.00)	8(100.0)	8(100.00)
Total	8(25.00)	8(25.00)	8(25.00)	8(25.00)	32(100.00)

TABLE X - CLASSIFICATION MATRIX FOR DATA SET TWO

	Predicted site membership and percentage in brackets set two				
Site	1	2	3	4	
1	11(91.67)	0(0.00)	1(8.33)	0(0.00)	12(100.00)
2	0(0.00)	7(58.33)	0(0.00)	5(41.67)	12(100.00)
3	0(0.00)	0(0.00)	12(100.00)	0(0.00)	12(100.00)
4	0(0.00)	4(33.33)	0(0.00)	8(66.67)	12(100.00)
Total	11(22.98)	11(22.98)	13(27.08)	13(27.08)	48(100.00)

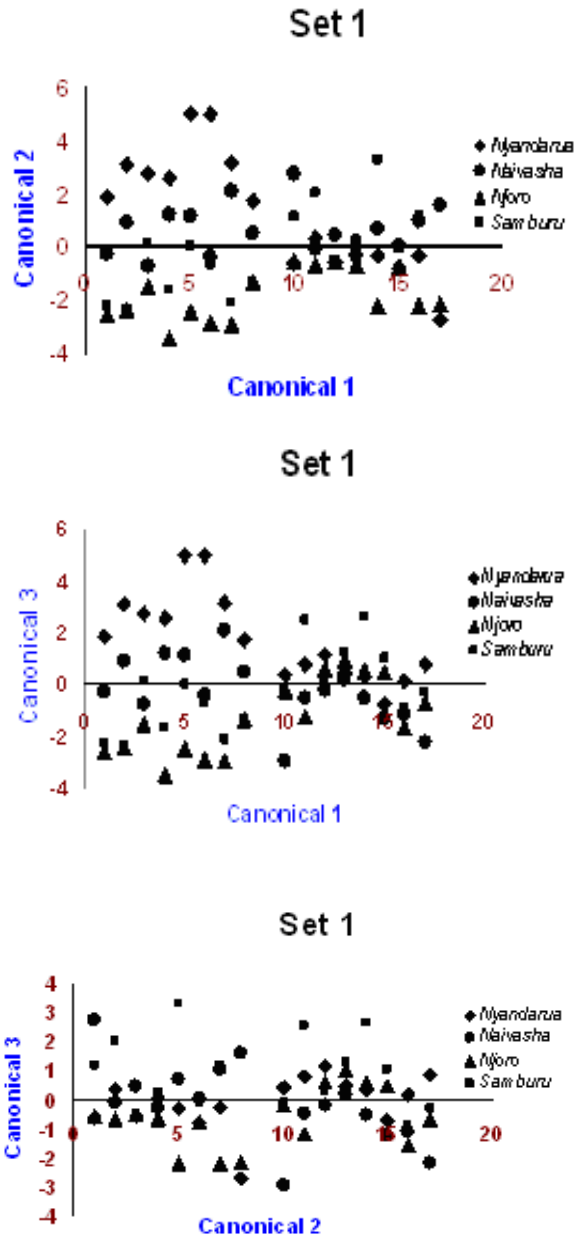

Figure 1. Plot of (a) Canonical1 vs Canonical2; (b) Canonical1 vs Canonical3; (c) Canonical2 vs Canonical3

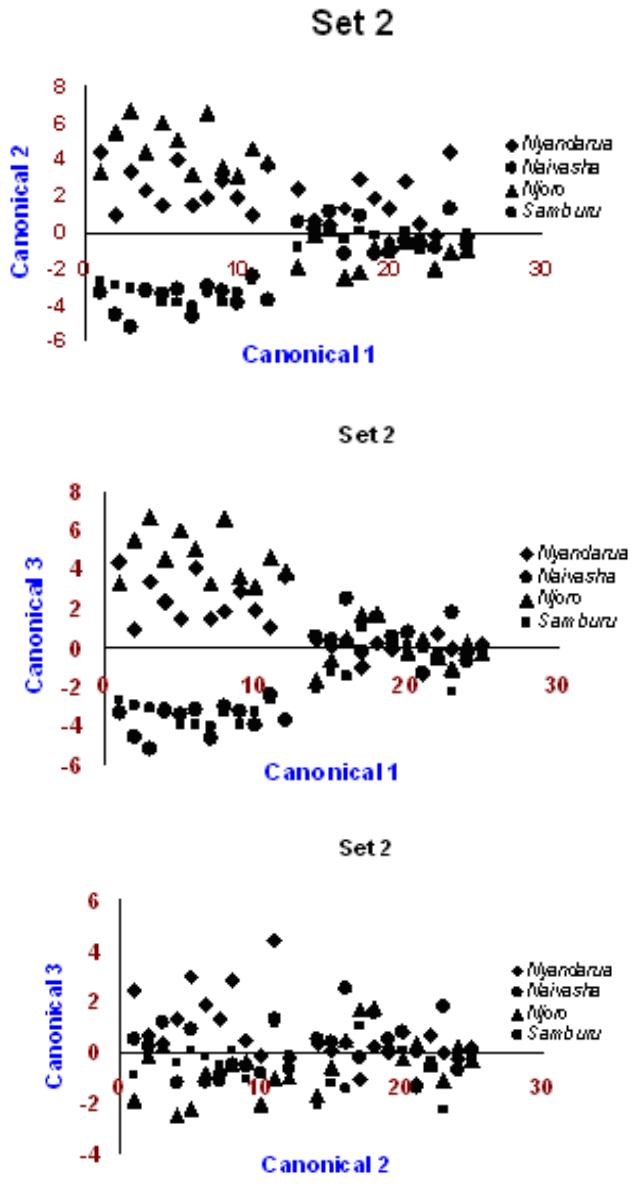

Figure 2. Plot of (a) Canonical1 vs Canonical2; (b) Canonical1 vs Canonical3 and (c) Canonical2 vs Canonical3

CONCLUSION

Stepwise discriminant analysis ranks the most useful characters for use in sites discrimination. The optimum variables used to produce canonical variables were determined. The scatterplots displaying canonical variables show all the individual observations; thus, any overlap between populations or sites can readily be examined. Stepwise and canonical analysis are recommended as useful tools for mineral discrimination.

ACKNOWLEDGEMENT

The authors wish to acknowledge the financial support by NCST and the KARI Director for provision of research facilities.

REFERENCE

[1] Van Gaastel, A.J.G (1978). Napier grass breeding at the national Agricultural Research Station Kitale Kenya. Ministry of Agriculture, Kenya breeding Project Technical Report GBP/AUG/78. p45

[2] Association of Official analytical chemistry (AOAC), (1990). Official method of analyses 15th Ed. AOAC, Washington, D.C USA.

[3] Bains, K.S., and K.C. Sood 1984. Resolution of genetic divergence for choice of parents in soybean breeding. *Crop Improvement* 11: 20-24.

[4] Boles, J.A. and J.E. Swan. 2002. Processing and sensory characteristics of coged roast beef: effect of breed, age, gender and storage conditions. *Meat Science*. 62: 419-427.

[5] Cruz-Castillo, J.G., S. Ganeshanandam, B.R. Lawes, C.R.O. Lawoko, and D.J. Woolley. 1994. Applications of canonical discriminant analysis in horticultural research. *HortScience.* 29: 1115-1119.

[6] Dillon, W.R., and M. Goldstein . 1984. Multivariate analysis methods and application. John Wiley and Sons, New York.

[7] Humphreys, M.O. 1991. A genetic approach to the multivariate differentiation of perennial ryegrass cultivars. *Heredity*. 66: 437-443.

[8] Kenkel, N.C., Derksen, D.A., Thomas, A.G. and Watson, P.R. 2002. Multivariate analysis in weed science research. *Weed Science*. 50: 281-292.

[9] Khattree, R. and D.N. 2000. Multivariate data reduction and discrimination with SAS software. SAS Inc. Cary, NC.

[10] Llames, C.R. and J. Fontaine (1994). Determination of amino acid in feeds: Collaborative study. *Journal of Association of official Analytical Chemistry. Intern*, 77: 1362 1402

[11] Lukibisi, F.B., J.N. Kariuki, T. Lanyasunya and D.M. Kuria (2008). Use of multivariate statistical analysis for on-farm livestock research data.

[12] Maria Loudes Gonzalez-Miret, Anass Terrab, Dolores Hernanz, Maria Angeles Fernandez-Recamales, and Francisco J. Heredia.. 2005. Multivariate Correlation between Color and Mineral Composition of Honeys and by Their Botanical Origin. *Journal of Agricultural and Food Chemistry*. 53: 2574-2580.

[13] McDonagh, M.B., Herd, R.M., Richardson, E.C., Olly, U.H., Archer, J.A., and Arthur, P.F. 2001. Meat quality and the calpain system of feedlot steers following a single generation of divergent selection for residual feed intake. *Australian Journal of Experiments in Agriculture*. 41: 1013-1021.

[14] Riggs, T.J. 1973. The use of canonical analysis for selection within a cultivar of spring barley. *Ann. Appl. Biol*. 74: 249-258.

[15] SAS Institute, Inc. 2003. SAS user's guide: Statistics. SAS Institute, Cary, NC.

[16] Vaylay, R.; and E van Santen (2002). Application of canonical discriminant analysis for the assessment of genetic variation in tall fescue. *Crop Science* 42: 534-539.