International Journal of Farm Sciences **5(3)** : 151-158, 2015

Statistical segregation of apple trees in relation to morphological characters using discriminant analysis

ANJU SHARMA, PK MAHAJAN, R MALHOTRA*, OK BELWAL
and SATISH K SHARMA**

Dr YS Parmar University of Horticulture and Forestry
Nauni, Solan 173230 HP, India
***National Dairy Research Institute, Karnal 132001 Haryana, India**
****HNB Garhwal University, Srinagar, Garhwal 246174 Uttarakhand, India**
Email for correspondence: anjusharma_uhf@yahoo.com

ABSTRACT

Apple is the most dominating fruit crop in Himachal Pradesh. It accounts for 48 per cent of the total area under fruit crops and more than 86 per cent of the total fruit production of the state. This study was focused on discriminant analysis for testing groups' differences with respect to a set of predictors or independent variables and to identify the predictor variables that best discriminated between two groups viz High Yielder and Low Yielder. It further involved the formulation of the categorization rule for allocating an apple tree to High Yielder and Low Yielder on the basis of different morphological characters of apple for both groups by developing a linear discriminating function rule to classify observations in the defined groups. The study suggested that for assigning a new tree with observed morphological characters (variables) to one of the groups the value of discriminant function or classification score (C_j) of both High Yielder and Low Yielder models is to be computed and the observation should be assigned to the group with highest score. Thus the rule is to allocate the tree to the group giving highest discriminant score. Error rate was also computed to evaluate the performance of the classification procedure. The discriminate function revealed a significant association between groups and all predictors accounting for 43.21 per cent between group variability although closer analysis of the structure matrix revealed yield (0.737) and number of flowers (0.448) to be two most important significant predictors with height and fruit weight as poor predictors. The cross validated classification showed that overall 81 per cent observations were correctly classified.

Keywords: Multivariate analysis; CDA; discriminant function; apple

INTRODUCTION

Apple is a predominant fruit crop of Himachal Pradesh and in recent years it is leading cash crop amongst fruit crops. It alone accounts for about 48 per cent of total area under fruit crops (211295 ha) and more than 86 per cent of the total fruit production (1027.82 thousand tons). The area under apple and its production has increased from 88673 hectares and 49.13 thousand tons in 1999-00 to 101485 hectares and 892.11 thousand tons in 2010-11 respectively (Anon 2012) . Shimla, Kullu, Mandi , Chamba and Kinnaur are apple producing districts of Himachal Pradesh. Shimla district of Himachal Pradesh being the leading apple producing district with 336753 MT which accounts for >50 per cent of total apple production of the state (Anon 2009) forms the basis for selecting it as the study area.

It is a common practice in horticultural experiments to categorize trees as low or high yielders on the basis of a single characteristic ie average fruit weight. This approach of categorization is statistically weak. The need was therefore felt to define a systematic and statistically valid procedure for categorization of trees as Low and High Yielders. Discriminant analysis (DA) is a multivariate technique based on linear combination of predictor variables and it provides the best discrimination point between the groups. Thus the main objective of the present study was to carry out DA by performing commonly used DA techniques available in SAS systems namely canonical and discriminant function analysis to find out the basis of two sets of observations from different populations a function which could best discriminate between the two populations namely High Yielder and Low Yielder and the relationship between the thirteen predictor variables and a grouping variable with an assumption that the variables are normally distributed in each of the groups and that the covariance matrices are the same (Fernandez 2002).

Betz (1987) has also described the method discriminant analysis by including the concept of discriminant function, discriminant score, group centroid, discriminant weights and loadings.

MATERIAL and METHODS

Data were recorded on different tree growth (morphological) characteristics viz X_1= age (years), X_2= girth (m), X_3= height (m), X_4= spread (m), X_5= volume (m^3), X_6= number of main branches, X_7= number of secondary branches, X_8= number of spurs per tertiary branch, X_9= length of spurs (cm), X_{10}= number of flowers per tertiary branch, X_{11}= number of fruits per tertiary branch, X_{12}= fruit weight (g) and X_{13}= yield of tree (kg) at three different locations of district Shimla of Himachal Pradesh namely Shimla, Kotkhai and Theog from a random sample of 300 apple trees from randomly selected orchard at each location.

For carrying out the discriminant analysis data thus collected from all (300 observations) the three locations were divided into two populations irrespective of age of trees namely High Yielder and Low Yielder on the basis of information on above referred characters from previous years' data. In the second step a discriminant function was fitted by considering the above referred 13 characteristics (predictors or discriminating variables) given by:

$$C_k = C_{k0} + C_{k1}X_1 + C_{k2}X_2 + ... + C_{km}X_m$$

where

C_k is the classification score or discriminant function for group k.
Cs are the discriminant coefficients for mth variable
Xs are observed values for mth variable
M is the number of predictor variables

Thus the study variables included 13 continuous numeric variables and one categorical or the grouping variable (Yielder) with two levels High Yielder and Low Yielder.

RESULTS and DISCUSSION

Population under consideration comprised of 149 (49.67%) high yielding (High Yielder) and 151 (50.33%) low yielding (Low Yielder) trees. Of these two groups the maximum number of discriminant functions produced was one. The results obtained from canonical linear discriminant analysis which is the classical form of discriminant analysis are as follows. Different multivariate statistics like Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace and Roy's Greatest Root were employed using SAS software to test the significance of discriminant function and the corresponding values obtained were 0.5679, 0.4320, 0.7606 and 0.7606 respectively with F-value equal to 16.73 for all statistics. The F-statistic for Roy's Greatest Root is an upper bound and for Wilks' Lambda it is exact. Wilks' Lambda statistics was used to test the null hypothesis (H_0) that there was no linear relationship between the predictors and the grouping variable. It was used to test which variable would contribute significance in discriminat function. Closer The Wilks' Lambda is to zero more the variable contributes to the discriminant function. The F-statistic was used to test the significance of Wilk's Lambda. Since the p-value was found to be less than 0.05 it can be concluded that the corresponding function explained the group membership well indicating thereby the significance of the discriminant function. It further provided the proportion of total variability not explained. Thus 56.79 per cent of variation in the grouping variable remained unexplained. Roy's Greatest Root gave the highest eigenvalue.

The analysis further showed that canonical correlation, adjusted canonical correlation and squared canonical correlation were found to be 0.657, 0.641 and 0.432 respectively with approximate standard error of 0.033. The canonical correlation of 0.657 suggested that the model explained 43.2 per cent of the variation in the grouping variable ie whether a tree is a High Yielder or Low Yielder. The eigenvalue provided information on discriminate function (equation) produced. The eigenvalue of the only discriminant function was obtained as 0.761. It revealed the canonical correlation for the discriminant function and the amount of variance shared the linear combination of variables. Likelihood ratio was found to be 0.568. The R^2 (0.432) between Can1 and the class variable as given by the squared canonical correlation was much higher than either univariate R^2.

The highly significant values of above statistics resulted in rejection of null hypothesis. Also the highly significant F and p values showed no evidence in favour of the hypothesis that the discriminant function had no discriminating ability or in other words discriminant function surely had the desired /tested ability.

Canonical structures and standardized canonical coefficients have been presented in Table 1.The total sample standardized canonical discriminant coefficients were used to rank the importance of each variable. The sign indicated the direction of the relationship. A high standardized discriminant function coefficient (2.069) indicated that the groups differed a lot on that X_{13} ie Yield variable. Volume (X_5) and number of flowers per tertiary branch (X_{10}) were next in importance as a predictor. These variables with large coefficients stand out as those that strongly predict allocation to High Yielder or Low Yielder groups. Other variables were less successful as predictors.

Canonical structures indicate the relative importance of the predictors. Total canonical structure indicated the positive correlation between the discriminant function and each variable. These Pearson coefficients are structure coefficients or discriminant loadings. They are like factor loadings of the variables on discriminant function. By identifying the largest loadings for each discriminate function the researcher gains insight into how to name each function. According to total canonical structure the variables like girth (X_2), number of spurs (X_8), flowers (X_{10}), fruits (X_{11}) and yield (X_{13}) were important and closely related to discriminant function as were above 0.30. From pooled within canonical structure it was observed that yield (X_{13}) had the highest correlation (0.64) with the discriminant function followed by number of flowers (X_{10}), number of spurs (X_8) and number of fruits (X_{12}) of apple trees population under study. Thus the first discriminator indicated that the groups differed with respect to the yield obtained from an apple tree.

The raw canonical coefficients or the unstandardized canonical coefficients for the first canonical variable (Can1) revealed that the classes differed most widely on the linear combination indicating the partial contribution of variables to the discriminate function controlling for all other variables in the equation:

Table 1. Canonical structure loadings and standardized canonical coefficients of variables under study

Variable	Canonical structure			Standardized canonical coefficient		
	Total canonical structure Can1	Between canonical structure Can1	Pooled within canonical structure Can1	Total-sample standardized canonical coefficient Can1	Pooled within-class standardized canonical coefficient Can1	Raw canonical coefficient Can1
X_1	0.234	1.000	0.179	-0.524	-0.518	-0.055
X_2	0.308	1.000	0.237	-0.361	-0.354	-1.366
X_3	0.069	1.000	0.052	-0.019	-0.019	-0.002
X_4	0.165	1.000	0.125	0.491	0.489	0.067
X_5	0.113	1.000	0.086	-1.041	-1.039	-0.000
X_6	0.238	1.000	0.182	-0.004	-0.004	-0.003
X_7	0.266	1.000	0.204	0.150	0.147	0.024
X_8	0.394	1.000	0.307	-0.439	-0.425	-0.084
X_9	0.156	1.000	0.118	-0.416	-0.414	-2.938
X_{10}	0.448	1.000	0.353	0.769	0.737	0.043
X_{11}	0.389	1.000	0.303	-0.189	-0.183	-0.019
X_{12}	0.084	1.000	0.064	0.004	0.004	0.000
X_{13}	0.737	1.000	0.635	2.069	1.813	0.020

$$D_j = -\ 0.05\ X_1j - 1.37\ X_2j + 0.07\ X_4j + 0.02\ X_7j - 0.08X_8j - 2.94\ X_9j + 0.04\ X_{10}j - 0.02X_{11}j + 0.02X_{13}j$$

where
D_j is the discriminant score for j^{th} observation and X_{ij} is the j^{th} observation for the i^{th} variable

A further way of interpreting discriminant analysis results is to describe each group in terms of its profile using the group means of the predictor variables. These group means are called centroids. Class means on canonical variables or the group centroids for High Yielder and Low Yielder (in Can1) groups were calculated as 0.875 and -0.863 respectively. The larger difference between the canonical group means indicated the better predictive power of the canonical discriminant function in classifying observations. Cases with scores near to a centroid are predicted as belonging to that group.

Linear discriminant function C_k also called ‘classification score’ for each observation for these groups has been presented below. These were the linear combinations of the variables that define each group. A constant or intercept term is also included.

High Yielder: C_k= $-127.97761 + 0.38060X_1 - 21.09591 X_2 + 0.55922X_3 + 0.84238X_4 - 0.00524X_5 - 0.00865X_6 + 0.97401X_7 + 5.28606X_8 + 161.27807X_9 - 1.12366X_{10} - 0.12579X_{11} + 0.11880X_{12} - 0.11426X_{13}$

Low Yielder: C_k= $-133.33814 + 0.47655X_1 - 18.72138X_2 + 0.56319X_3 + 0.72529X_4 - 0.00450X_5 - 0.00309X_6 + 0.93157X_7 + 5.43258X_8 + 166.38671X_9 - 1.19791X_{10} - 0.09190X_{11} + 0.11857X_{12} - 0.14904X_{13}$

where C_k is the classification score for group k.

For one observation its score can be computed for each group by substituting the values of variables according to equations mentioned above. The observation should be assigned to the group with highest score. The coefficients are helpful in deciding which variable affects more in classification.

Table 2 shows number of observations and per cent classified in re-substitution and cross validation summary into two groups of High Yielders and low Yielders along with error count estimates viz the prior probability of each group and the rate of mis-classification. The rows in the resubstitution summary table (also called as classification count table) represented the observed groups of the observations (ie Total High Yielder= 149 and Total Low Yielder= 151) and the columns represented the predicted groups (ie Total High Yielder= 128 and Total Low Yielder= 172). Values in the diagonal of the table reflected the correct classification of observations into groups. When prediction is perfect all cases will lie on the diagonal. It is evident from the Table that 74.50 per cent (111) High Yielders observations were correctly classified leaving a 25.50 per cent rate of error count. Low Yielders did better with 11.26 per cent error rate. The overall rate of error count was 18.33 per cent.

The cross validation is often termed a 'jack-knife classification' in that it successively classifies all cases but one to develop a discriminant function and then categorizes the case that is left out. This process is repeated with each case left out in turn. This cross validation produces a more reliable function. The results of cross validation revealed that High Yielders exhibited an error rate of 26.17 per cent while 11.92 per cent rate of error count was shown by the Low Yielders. Overall 19 per cent of the observations were mis-classified.

Table 2. Re-substitution and cross validation of trees into high yielders and low yielders

	Re-substitution summary using discriminant function			Cross validation summary using discriminant function		
From yielder	High yielder	Low yielder	Total	High yielder	Low yielder	Total
High yielder	111 (74.50)	38 (25.50)	149 (100.00)	110 (73.83)	39 (26.17)	149 (100.00)
Low yielder	17 (11.26)	134 (88.74)	151 (100.00)	18 (11.92)	133 (88.08)	151 (100.00)
Total	128 (42.67)	172 (57.33)	300 (100.00)	128 (42.67)	172 (57.33)	300 (100.00)
Priors	0.49667	0.50333		0.49667	0.50333	
Rate of error count	0.2550	0.1126	0.1833	0.2617	0.1192	0.1900
Rate of priors	0.4967	0.5033		0.4967	0.5033	

Figures in parentheses indicate the per cent classified into groups

Soti Negrin (1970) applied discriminant analysis to experimental data on sugarcane. Multiple group discriminant function approach was carried out by Havaldar and Sharma (1997) to classify the yields of crops in terms of soil characteristics. Four groups in the form of cross classification of two levels of nitrogen (low-high) with two levels of yield (low-high) were made in revealing the relative importance of soil characteristics of practical utility. In a study conducted by Mangas et al (1999) several pattern recognition procedures including principal component analysis (PCA), linear discriminant analysis (LDA) and partial least squares (PLS-1) were applied to the data on phenolic compounds in 46 Spanish cider apple varieties in an attempt to classify the samples into bitter and non-bitter categories. Reliable decision rules were obtained by both LDA and PLS-1. LDA achieved 91.3 and 85.7 per cent correct classification respectively for internal and external evaluation of the model. Mahajan et al (2008) suggested use of discriminant and principal component analyses in apricot crop. They proposed a rule to allocate the tree to High Yielder group if D was greater than or equal to m where m is the value of cut-off point (= 0.359) and D is the discriminant function. Ballabio et al (2012) studied the relationships between apple texture and rheological parameters by means of multivariate analysis. Principal component analysis enhanced the interpretation of the relationships between crispness and rheological variables while linear discriminant analysis showed a clear trend of separation among the different apple cultivars according to their crispness score. Finally the new apple populations were projected in the multivariate model in order to predict their crispness.

CONCLUSION

The discriminant analysis indicated that the groups differed a lot on the variable X_{13} ie yield (obtained from an apple tree) having the highest correlation (0.64) with the discriminant function followed by number of flowers (X_{10}), number of spurs (X_8) and number of fruits (X_{12}) of apple tree population under study. The larger difference between the canonical group means indicated the better predictive power of the canonical discriminant function in classifying observations. Linear discriminant or classification function score for High Yielder and Low Yielder groups can be calculated by substituting the values of variables in the equations obtained. The observation should be assigned to the group with highest score. Yield and number of flowers were found to be two most important predictors while height and fruit weight were poor predictors. The cross validated classification showed that overall 19 per cent observations were not correctly classified.

REFERENCES

Anonymous 2009. Administrative report. Directorate of Horticulture, HP, India.

Anonymous 2012. Statistical outline of Himachal Pradesh 2011-12. Department of Economics and Statistics, Shimla, Himachal Pradesh, India.

Ballabio D, Consonni V, Costa . 2012. Relationships between apple texture and rheological parameters by means of multivariate analysis. Chemometrics and Intelligent Laboratory Systems **111(1):** 28-33.

Betz NE 1987. Use of discriminant analysis in counselling psychology research. Journal of Counseling Psychology **34(4):** 393-403.

Fernandez GCJ 2002. Discriminant analysis, a powerful classification technique in data mining. Proceedings, 27th Annual SAS Users Group International Conference, Statistics and Data Analysis, pp 247-256.

Havaldar YN and Sharma KMS 1997. A multiple discriminant function approach to study the crop response to soil characters (Bengal gram crop). Karnataka Journal of Agricultural Sciences **10(1):** 124-128.

Mahajan PK, Thakur M and Negi YS 2008. Use of discriminant and principal component analyses in horticultural research– a case of apricot crop. Horticulture Journal **21(1):** 7-10.

Mangas JJ, Rodríguez R, Suárez B, Picinelli A and Dapena E 1999. Study of the phenolic profile of cider apple cultivars at maturity by multivariate techniques. Journal of Agricultural and Food Chemistry **47(10):** 4046-52.

Soti Negrin E 1970. Application of discriminant analysis to experimental data on sugar cane. Agronomia Tropical (Maracay) **20:** 5-14.

Received: 11.9.2014 ***Accepted: 23.11.2014***