Discriminant and Classification Analysis of Health Status of Bell Pepper (*Capsium annum*)

Oludare S. Ariyo, O. Arogundade and Ayoola M. Abdul-Rafiu
National Horticultural Research Institute, PMB, 5432, Jericho Idi-Ishin, Ibadan, Nigeria

**Abstract:** This study is considered as an attempt for employing the discriminant analysis and classification method for the purpose of achieving the assessment of discriminant function through which we can discover the health status of bell pepper based on measured growth parameters. Linear Discriminant Function (LDF) was used as a tool for statistics analysis. It was estimated on the bases of data on growth attributes for data taken from 6 plants from each plot; 3 healthy and 3 diseased. The growth attributes taken were Plant height, number of leaves and number of fruit six and ten weeks after transplanting. The study results show high significant difference between both groups based on Hotelling T² and wilks’ lambda statistics and different variance covariance structure based on Berlet’s approximation to Chi-square statistics. Finally, the study results showed that the adapted classification rule lead to the classification of about 83% of the bell pepper to the group which they belong and 17% to the wrong group. These results shows that discriminant analysis may be an efficient and useful tool for predicting the health status of bell pepper on the bases of a few growths attributes that are routinely assessed by researchers and farmers.

**Key words:** Hotelling T², Linear Discriminant Function (LDF), wilks’ lambda

**INTRODUCTION**

Capsicum pepper refer primarily to *Capsicum annum* L and *Capsicum frutescens* L., plants used in the manufacture of selected commercial products known for their pungency and colour (Simon *et al*., 1984). Capsium pepper is of the night shade family solanaceae along with egg plant and tomato, it is a vegetable grown mainly for its fruit (Hedge, 1997). Tropical south America, particular Brazil is thought to be the original home of pepper, it is now widely cultuvavated in America , South America, Europe, some counties in Asia pacific and African (Shoemaker and Tesky, 1995).

Discriminant Analysis (DA) encompasses procedures for classifying observation into groups (i.e., Predictive discriminant analysis) and describing the relative importance of variable for distinguishing amongst groups (Lix and Sojobi, 2010). It based on the conception of a function which could provide the identification of the subject of analysis (Kara *et al*., 2005). Discriminant analysis could be put into application in case data matrix has a common variance-covariance matrix (Atewi *et al*., 2006; Mikkonen *et al*., 2006). Tai and Pan (2007) proposed modified linear discriminant methods to integrate biological knowledge of gene functions (or variable groups) into classification of microarray data. Discriminant analysis methods were integrated to food authentication (Murphy *et al*., 2010) using a model-based discriminant analysis methods that includes variable selection. The DA was applied to study the effect of Irrigation Frequency on the growth and yield of Okro in Ibadan (Dauda *et al*., 2007). In recent years, a number of developments have occurred in DA procedures for the analysis of data from repeated measures designs. Specifically, DA procedures have been developed for repeated measures data characterized by missing observations and/or unbalanced measurement occasions, as well as high-dimensional data in which measurements are collected repeatedly on two or more variables (Lix and Sojobi, 2010). Yaakob *et al*. (2010) used DA to exploit the classification of Lard and other commercial vegetable oil and animal facts, the result shows that all vegetable fats/oils and animal facts, including lard are clustered in a district group. Discriminant analysis has been used successfully in horticulture in classifying plant material (Cruz-Castillo *et al*., 1994; Ebdon *et al*., 1998; Pydipati *et al*., 2006). Few literatures has applied DA to determine the health Status of an horticultural crop due to its complex nature, this paper therefore considered as an attempt for employing the discriminant analysis and
classification method for the purpose of achieving the assessment of discriminant function through which we can discover the health status of bell pepper based on measured growth parameters.

**THEORETICAL SECTION**

Another approach to the discriminnat problem based on a data matrix can be made by not assuming any particular parametric form for the distribution of the population \(\pi_1, \pi_2, ..., \pi_n\) but by merely looking for a rule to discrimante between them (Mardia et al., 1971). Fisher (1938) looks at the linear function \(a'x\) which maximized the ratio of the between-group sum of squares to the within-groups sum of squares that are a linear combination of the columns of \(X\). Then \(y\) has the total sum of squares:

\[
Y'Hy = a'XHAa = a'TA
\]

which can be partitioned as a sum of within-groups sum of squares:

\[
\sum y'_iH_iy_j = \sum y'_iX'_iH_iX_ja = a'Wa
\]

Plus the between-groups sum of squares:

\[
\sum n_i(\bar{y}_i - \bar{y})^2 - \sum n_i\left[a'(\bar{x}_i - \bar{x})\right]^2 = a'Ba
\]

Where the mean of the ith is sub-vector of \(y\) and \(x\) is the (centring matrix). The ratio is given by:

\[
\frac{a'Ba}{a'Wa}
\]

If is the vector which maximized (4) we shall call the linear function \(a'x\) Fisher’s discriminant function or the first canonical variant. Fisher’s discriminant function is most important in the special case of \(g = 2\) groups. Then \(B\) has rank one and can be written as:

\[
B = \left(\frac{n_1n_2}{n}\right)dd'
\]

where \(d = \bar{x}_1 - \bar{x}_2\). This \(W^{-1}B\) has only one non-zero eigenvalue which can be found explicitly. This eigenvalue equals:

\[
\text{tr}W^{-1}B = \left(\frac{n_1n_2}{n}\right)d'W^{-1}d
\]

And corresponding eigenvector i:

\[
a = W^{-1}d
\]

The discriminant rule becomes: allocate \(x\) to \(\pi_1\) if \(>0\) to \(\pi_2\) otherwise. Testing the significant of the separation between the two populations.

Suppose the population \(\pi_1\) and \(\pi_2\) are multivariate with a common covariance matrix \(\Sigma\), we know that to test the following statistical hypothesis:

\[
H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 \neq \mu_2
\]

we can use the hotelling’s \(T^2\)-statistics as:

\[
T^2 = \frac{n_1n_2}{n_1 + n_2}\left(\bar{X}_1 - \bar{X}_2\right)'S_{pooled}^{-1}\left(\bar{X}_1 - \bar{X}_2\right)
\]

(Anderon, 1984) (9)

\[
T^2 = \frac{(n_1 + n_2 - 2)}{(n_1 + n_2 - p - 1)}F_{p,n_1+n_2-p-1}(\alpha)
\]

\[
F = \frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - 2)}T^2 \sim F_{p,n_1+n_2-p-1}
\]

where \(F_{p,n_1+n_2-p-1}(\alpha)\) is the value of \(F\) with degree of freedom of numerator \(p\) and degree of freedom of denominator \(n_1 + n_2 - p - 1\). So, we reject the null hypothesis \(H_0 : \mu_1 = \mu_2\) at: the level of significant \(a\) when:

\[
F > F_{p,n_1+n_2-p-1}(\alpha)
\]

(12)

Apparent error rate (APER) can be calculated from the confusion matrix, which shows actual versus predicted group membership. According to (Richard and Dean, 1998), the confusion matrix has the form:

\[
\begin{array}{cc}
\pi_1 & \pi_2 \\
\n_1c & n_{1M} = n_1 - n_{1c} \\
n_{2M} = n_1 - n_{1c} & n_{2c}
\end{array}
\]
where

\[ n_{1c} = \text{Number of } \pi_1 \text{ items correctly as } \pi_1 \text{ items} \]
\[ n_{2c} = \text{Number of } \pi_2 \text{ items correctly as } \pi_2 \text{ items} \]
\[ n_{1M} = \text{Number of } \pi_1 \text{ items misclassified as } \pi_2 \text{ items} \]
\[ n_{1M} = \text{Number of } \pi_2 \text{ items misclassified as } \pi_1 \text{ items} \]

The apparent error rate is then:

\[ \text{APER} = \frac{n_{1M} + n_{2M}}{n_1c + n_2c} \quad (13) \]

### APPLICATION SECTION

The experiment was carried out at the research farm of National Horticultural Research Institutes (NIHORT), Ibadan. The experiment was laid out in the Randomized Complete Block Design (RCBD) with three replications. The experiment field was first divided into three blocks. Each block was further divided into 10 unit plots. The size of each unit plot was 2 m × 2 m. The blocks and the plots were spaced at 0.5 m between and within. Thirty-five-day-old healthy seedlings were transplanted in the experimental plots at a spacing of 50 cm × 50 cm. The data taken were 6 plants from each plot; 3 healthy and 3 diseased. The diseased plants were showing varying symptoms of disease while the healthy plant was apparently healthy. The following growth attributes were taken Plant height, number of leaves and number of fruit. The recorded data were analysed using Statistical Analysis System (SAS) version 9.1.

### RESULTS AND DISCUSSION

The Shapiro-Wilks (W) test was used to test the normality assumption of the growth attributes of *Capsicum annum* across the health status. The normality assumption was significant at 5% level of significance; this implies that all the growth attributes are approximality normal (Table 1).

To test whether the population mean and differ significantly, Wilks’ lambda statistics (Table 2) was used. We conclude that the separation between the two groups (disease and Healthy *Capsicum annum*) is significant. This gives a logic justification for the importance of the discriminant function for discriminating or classifying *Capsicum annum* into diseased or healthy group. The value of Wilks’ lambda statistic is 0.3707 (Table 2), the complement of the Wilks’ lambda 1 - \( \Lambda \), can be used as a generalized \( R^2 \) indicator (Huberty and Olejnik, 2006) which indicates the amount of variance in the p-variables system that is attributed to the effect of the grouping variable. Therefore, about 60% of the growth attributes of *Capsicum annum* attributed to the diseased or Healthy *Capsicum annum*.

To test whether the variance covariance matrix of the two groups differs, Berlet’s approximation to chi-squares test was used, this gave an insignificant result, we conclude that the two group had the same variance covariance matrix, this justified the used of linear discriminant function.

The structure of Correlation between eight growth attributes (dependent variables) and the Linear Discriminant Functions (LDFs) are conveniently summarized in Table 3. This shows that the first LDF is labelled CAN1 and the second LDF CAN2. As seen from Table 3, X2 (Plant height at 10 week after transplant) has largest loading on the first LDF, followed by X8 (stem girth at 10 week after transplant), X1 (Plant height at 6

### Table 1: Test for normality (Shapiro-Wilk) for each of the growth attributes of bell pepper (*Capsicum annum*)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Statistics</th>
<th>p-value</th>
<th>Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>0.8965</td>
<td>Pr&gt;W</td>
<td>0.0013</td>
</tr>
<tr>
<td>X2</td>
<td>0.9252</td>
<td>Pr&gt;W</td>
<td>0.0001</td>
</tr>
<tr>
<td>X3</td>
<td>0.8019</td>
<td>Pr&gt;W</td>
<td>0.0001</td>
</tr>
<tr>
<td>X4</td>
<td>0.79115</td>
<td>Pr&gt;W</td>
<td>0.0001</td>
</tr>
<tr>
<td>X5</td>
<td>0.8170</td>
<td>Pr&gt;W</td>
<td>0.0001</td>
</tr>
<tr>
<td>X6</td>
<td>0.8683</td>
<td>Pr&gt;W</td>
<td>0.0001</td>
</tr>
<tr>
<td>X7</td>
<td>0.9351</td>
<td>Pr&gt;W</td>
<td>0.0003</td>
</tr>
<tr>
<td>X8</td>
<td>0.9366</td>
<td>Pr&gt;W</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

### Table 2: Wilks’ Lambda and chi-square test values pertaining to discriminant function

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Hypothesis</th>
<th>value</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks’ Lambda</td>
<td>H0 : ( \mu_1 = \mu_2 )</td>
<td>0.3707</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>H1 : ( \mu_1 \neq \mu_2 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>H0 : ( \Sigma_1 = \Sigma_2 )</td>
<td>186.370</td>
<td>0.9927</td>
</tr>
<tr>
<td></td>
<td>H1 : ( \Sigma_1 \neq \Sigma_2 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 3: Pooled within canonical structure

<table>
<thead>
<tr>
<th>Variable</th>
<th>CAN1</th>
<th>CAN2</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>-0.604467</td>
<td>0.438617</td>
</tr>
<tr>
<td>X2</td>
<td>0.971020</td>
<td>0.458473</td>
</tr>
<tr>
<td>X3</td>
<td>0.242800</td>
<td>0.362540</td>
</tr>
<tr>
<td>X4</td>
<td>-0.357270</td>
<td>0.246961</td>
</tr>
<tr>
<td>X5</td>
<td>-0.418430</td>
<td>0.197429</td>
</tr>
<tr>
<td>X6</td>
<td>0.575110</td>
<td>0.165260</td>
</tr>
<tr>
<td>X7</td>
<td>0.316336</td>
<td>0.576054</td>
</tr>
<tr>
<td>X8</td>
<td>0.791720</td>
<td>0.702272</td>
</tr>
</tbody>
</table>
Table 4: Coefficient of Linear Discriminant Function (LDF) status in health status of *Capsicum annum*

<table>
<thead>
<tr>
<th>Variables</th>
<th>Diseased</th>
<th>Health</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-22.92579</td>
<td>-39.52036</td>
</tr>
<tr>
<td>X1</td>
<td>0.07742</td>
<td>-0.14573</td>
</tr>
<tr>
<td>X2</td>
<td>0.39047</td>
<td>0.06337</td>
</tr>
<tr>
<td>X3</td>
<td>0.31987</td>
<td>0.37266</td>
</tr>
<tr>
<td>X4</td>
<td>-0.15112</td>
<td>-2.65213</td>
</tr>
<tr>
<td>X5</td>
<td>-2.01418</td>
<td>-2.65213</td>
</tr>
<tr>
<td>X6</td>
<td>1.49387</td>
<td>2.11115</td>
</tr>
<tr>
<td>X7</td>
<td>41.2201</td>
<td>48.48411</td>
</tr>
<tr>
<td>X8</td>
<td>35.6086</td>
<td>50.0531</td>
</tr>
</tbody>
</table>

Table 5: Summary of classification with cross-validation using growth attributes as prediction healthy status in bell pepper (*Capsicum annum*)

<table>
<thead>
<tr>
<th>From</th>
<th>Diseased</th>
<th>Healthy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diseased</td>
<td>27(90%)</td>
<td>2(10%)</td>
<td>30(100%)</td>
</tr>
<tr>
<td>Healthy</td>
<td>3(6.67%)</td>
<td>28(93.33%)</td>
<td>30(100%)</td>
</tr>
<tr>
<td>Total</td>
<td>30(50%)</td>
<td>30(50%)</td>
<td>60(100%)</td>
</tr>
</tbody>
</table>

The linear discriminant analysis coefficient in Table 4 was used to predict, using Cross-validation error rate, the number of healthy *Capsicum annum* that are correctly classified. The summary is presented in Table 5, 90% of diseased *Capsicum annum* was correctly classified while about 7% was misclassified. The Apparent error rate (APER), from Eq. (13), for the discriminant analysis is 83.3%. The adopted classification rule leads to about 83.3% of *Capsicum annum* correctly classified according to their healthy status which leads to about 17% of the pepper classified to the group which they do not belong to.

**CONCLUSION**

This study is considered as an attempt of using the discriminant and classification methods for the purpose of achieving a discriminant function, though which we can identify the healthy from the diseased bell pepper with few growth attributes which can easily measured by farmers.

In this study, Linear Discriminant Function (LDF) was used as a statistical tool for statistical analysis. The discriminant analysis adopted was found to effective (with about 83%) to classified bell pepper to its true health statuses. These results shows that discriminant analysis may be an efficient and useful tool for predicting the health status of bell pepper on the bases of a few growths attributes that are routinely assessed by breeders, researchers and farmers.

**REFERENCES**


