*Original Research Article*

# Yield estimation of Dolichos bean (Lablab purpureus L.) using different clustering techniques


Abstract

Dolichos bean (*Lablab purpureus* L.) is an important leguminous vegetable crop. Diversity of genes to a high extent in the gene pool provides better opportunities to develop appropriate plant varieties. Hence, finding advantageous traits or their combinations is the key step. Genetic diversity is important for parents to be chosen for potential breeding programmes. In general, genetically divergent parents are used in segregating generations to produce the optimal recombinants. Green seed yield of the crop is influenced by number of yield contributing characters such as, green pod yield, pods per plant, days to 50 percent flowering, number of primary branches, raceme per plant, nodes per raceme, raceme length and 100 fresh seed weight. The association between the characters and yield provides basis for further breeding programme. Cluster analysis was used to classify the germplasm accessions into different clusters based on the morphological traits of the accessions. The data on germplasm accessions (64) was collected from the Department of Genetics and Plant Breeding, UAS, GKVK, Bengaluru, Karnataka. Euclidean distance metric was used to measure the difference between germplasms. Elbow method indicated three optimum genotype clusters. Three clustering methods were used *viz.,* single, complete and average linkage. Dunn and Silhouette indices were used to validate the clusters. It was found that Average linkage method was best among the three. Further, in the third cluster by average linkage method with 12 genotypes was found to have highest green seed yield and green pod yield. Hence these 12 genotypes can be used for further breeding programmes.



**Key words:** Germplasm accessions, Green seed yield, Hierarchical method of clustering, Dendrogram, Single linkage, Complete linkage, Average linkage, Dunn and Silhouette indices.




## 1. Introduction

Dolichos Bean (*Lablab purpureus* L.)(2n=22) is a species of bean in the family Fabaceae. It is native to Africa and has been grown in the tropics for food. The plant is variable due to extensive breeding in cultivation.The genetically divergent parents usually are used in segregating generations to produce the optimal recombinants. Mating between genetically different parents is likely to produce a strong heterotic effect, and the crosses within the same species involving distant related parents reflect a huge spectrum of variations.

The correlation between different morphological parameters and related traits with yield gives an idea about attributes of growth and yield. Cluster analysis is technique used to classify the different objects into groups in such a way that the similarity between the objects which are maximal if they are belong to same group and minimal otherwise. The coefficient of correlation

can be used as the element of primary and secondary influence. Investigation of correlation and cluster analysis will help in understanding the extent of interconnection in both seed yield and their constituent characters.

## 2. Methods

The secondary data on yield and yield attributing characters comprising 64 accessions of Dolichos bean was collected from the Department of Genetics and Plant Breeding, UAS, GKVK, Bengaluru,as the experiment was conducted during the Kharif seasons of 2013, 2014 and 2015 were grown for the experimental purpose. The traits such as Days to 50 per cent flowering, Primary branches, Racemes per plant, Raceme length, Nodes per raceme, Pods per plant, Green pod Yield, Green seed yield and 100 fresh seed weight were included for the study to know how these factors had an influence on the green seed yield.

Cluster Analysis is an exploratory technique that seeks to classify data structures. Cluster analysis involves examining multivariate data and grouping objects into clusters. These clusters are constructed in such a way that items within a cluster are homogeneous and items between clusters are highly dissimilar or heterogeneous (Richard and Dean, 1998).
For this study method of Hierarchical was employed for clustering technique (Tiwari and Misra 2011).

### 2.1 Hierarchical method of clustering

This continues either through a series of mergers or through a series of successive divisions. Agglomerative hierarchical method of clustering starts with the individual objects, there will be as many clusters as objects were present. First, the most similar objects are grouped, and these initial groups are merged based on their similarities. Eventually, all subgroups are fused into a single cluster based on the similarity.

In divisive hierarchical approach,the original single group of objects is split into two subgroups, further split into subgroups which are dissimilar. The process continues until as many subgroups as objects, i.e. until each object forms a group, are present. Both agglomerative and divisive process findings can be represented in the form of a two-dimensional diagram known as Dendrogram.

**2.1.1 Single Linkage:** Groups are formed by merging nearest neighbors, *i.e.,* the smallest distance or the largest similarities, from the individual entities.

Initially, to get cluster (UV), we must find the smallest distance in D = $\{d_{ik}\}$ and merge the corresponding objects, say, U and V. The distance between (UV) and any other cluster W is determined by formula for step 3 of the general algorithm.

$D_{(uv)w} = \min \{d_{uv}, d_{vw}\}$

Single linkage clustering effects may be represented graphically in the form of a Dendrogram or Tree Diagram. In the tree, the branches represent clusters. At nodes whose locations along a distance (or similarity) axis indicate the degree at which the fusion occurs *i.e.,* the branches come together.

**2.1.2 Complete Linkage:** Here at each stage, the distance (similarity) between the clusters between the two components is defined by the distance (similarity), one that is most distant from each cluster. Thus, complete linkage ensures that all items in a cluster are with in some maximum distance (or minimum similarity) of each other.

The general agglomerative algorithm again starts by finding the minimum entry in D = $\{d_{ik}\}$ and merging the corresponding objects, such as U and V, to get cluster (UV). Step (iii) of general algorithm, the detachment amid (UV) and any other cluster W is

$D_{(uv)w} = \max \{d_{uv}, d_{vw}\}$

Here $d_{uw}$ and $d_{vw}$ are the distances between the most distant members of clusters U and W and clusters V and W.

**2.1.3 Average Linkage:** Average linkage views the distances between two clusters as the average distance between all pairs of objects where each cluster belongs to one member of a group. Again, distances or similarities may be the input to an average linkage algorithm, and the process may be used to group objects or variables. The typical linkage algorithm operates like the general algorithm, we start by searching the distance matrix D = $\{d_{ik}\}$ to find the nearest (most similar) objects like U and V. These objects are merged into the cluster (UV). The distance between (UV) and other cluster W is determined by step 3 of the general agglomerative algorithm

$$d_{(uv)w} = \frac{\sum_i \sum_j d_{ik}}{N_{(uv)} * N_w}$$

Where $d_{ik}$is the distance between object i in the cluster (UV) and object k in the cluster W, and where $d_{ik}$ is the member of objects in the clusters (UV) and W.

**2.2. Elbowmethod:** Elbow method is a method that looks at the percentage of variance as a function of the number of clusters explained. It helps in finding the optimum number of clusters using within sum of squares for the clusters.

**2.3. Dendrogram**

Dendrogram is also called a hierarchical tree diagram or map, which displays the relative size of the coefficients of proximity where the cases are combined. Larger the coefficient of distance, or the smaller the coefficient of similarity, the more clustering requires merging unlike individuals, which might not be ideal. Trees are usually represented horizontally, with each row representing a case on the Y axis, while the X axis is a rescaled representation of the coefficients of proximity. Small distance or high similarity cases are identical to each other.

Cases showing a low distance are similar, with a line joining them from the left of the Dendrogram, suggesting that they are agglomerated at a low distance coefficient into a cluster, suggesting that they are alike.

Both agglomerative and divisive Hierarchical clustering devices have been utilized in data analysis for the present study.

### 2.4. Cluster validation methods

Validation of the clusters was carried by using two measures *i. e.,* Dunn index and silhouette coefficient.

#### 2.4.1. Dunn index (>0):

It is the ratio of minimum of this pair wise distance as the inter-cluster separation (minimum separation) to the maximal intra-cluster distance (maximum diameter).

The Dunn index is another internal clustering validation measure which can be computed as follow:

$$D= \frac{Min.\ seperation}{Max.\ diameter}$$

#### 2.4.2. Silhouette coefficient (-1 to 1):

The silhouette analysis measures how well an observation is clustered and it estimates the average distance between clusters.

$$S_i=\frac{b_i-a_i}{\max(a_i,b_i)}$$

Where,

$a_i$ –average dissimilarity between i and all other points of the cluster to which i belongs.

$b_i$–dissimilarity between i and its “neighbor” cluster, i.e., the nearest one to which it does not belong.

## 3. Result and discussion

### 3.1. Optimum number of clusters:

Among several methods used to know the optimal number of clusters, Elbow’s approach has been found to be the best one. Using this approach for the analysis, it was determined that the optimum number of clusters can be 3. Figure 1 uses the Elbow test to determine the optimal number of clusters.

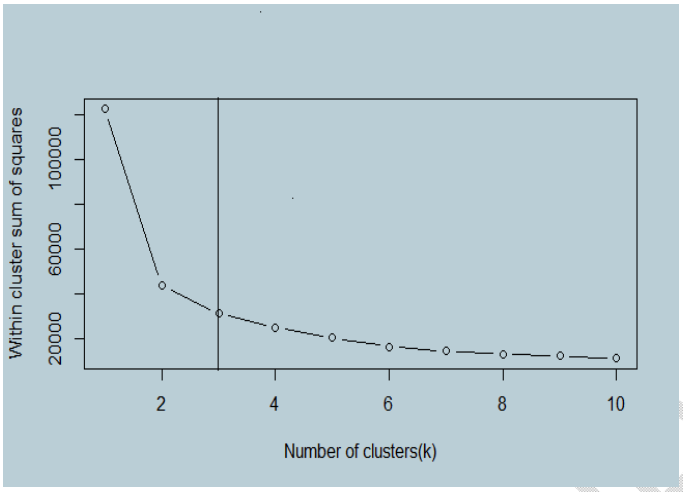


**Fig.1 Graph showing optimum number of clusters using Elbow's method**

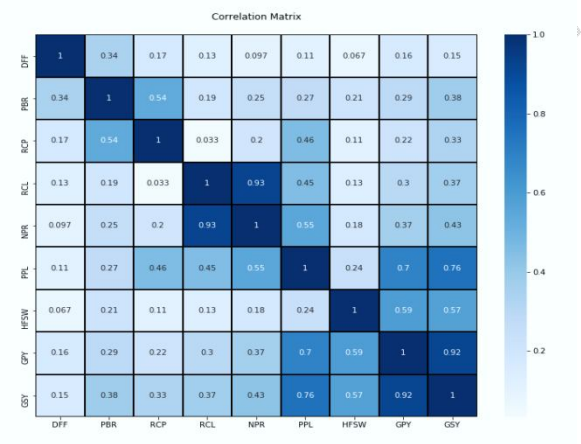


**Fig.2: Correlation plot for Green seed yield and the component traits of Dolichos bean germplasm accessions**

### 3.2. Correlation between morphological traits and Green seed yield:

Correlation between Green seed yield and other morphological traits is important since they indicate the contribution of those characters to the crop yield (Parmar *et al.*2013). Because the clusters are formed based on all the characters being taken into consideration. Figure 2 shows the correlation among all the characters. The figure illustrates that there is a high positive correlation (0.92) between green seed yield and green pod yield (Kousar*et al.*2007). Pods per plant character was also having high positive correlation

(0.76) with green seed yield. And all the other characters were having positive correlation with the green seed yield, with none having a negative correlation (Babu*et al*. 2012).

### 3.3.1. Single linkage method

Characters for the clusters obtained by employing single linkage method are listed in Table 1.It can be observed that the 1st cluster contains 61 germplasm accessions while the 2nd cluster had 2 germplasm accessions and the 3rd cluster had just one accession.It can be deduced by looking at table that the cluster diameter was estimated to be highest (662.138) for cluster 1. Likewise, for cluster 1, cluster wise within cluster average distances was found to be the largest (184.760). Cluster wise minimum distances of a point in the cluster to a point of another cluster were the same for both cluster 1 and cluster 2, with the value 46.975, however cluster 3 was having highest value (49.170) for this character. Cluster wise average distances of a point in the cluster to the points of other clusters is found to be highest for cluster 2 that is 240.371.Dendrogram for the corresponding table is as shown in Figure 3

### 3.3.2. Complete linkage method

In this approach the distance (dissimilarity) between clusters is determined by the distance between the two elements (dissimilarity), one from each cluster that is most distant. Complete linkage thus ensures that all objects in a cluster are within a certain maximum distance (or minimum similarity) of each other. The results of complete linkage method are represented in the Table 2 and the dendrogram for the corresponding table is as shown in Figure 4. Table 2 will make it evident that there are 11 germplasm accessions in the 1st cluster, 29 accessions in the 2nd cluster and 24 accessions are there in the 3rd cluster.

Characters assigned to the three clusters obtained using the complete linkage method are described in Table 2. From this table it was evident that the cluster diameter was highest (271.404) for cluster 3. Likewise, cluster wise within cluster average distances was found to be highest (120.020) for cluster 3. Cluster wise minimum distances of a point in the cluster to a point of another cluster are the same for both cluster 2 and cluster 3, with the value 14.300, while cluster 1 having highest value (19.740). Cluster 1 with a value of 291.063 for cluster wise average distances of a point in the cluster to the points of other clusters was found to be the largest.

**Table 1: Description of the cluster wise characters for Single linkage method**

| Sl. no. | Characters of individual clusters | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|---|
| 1 | Cluster size | 61 | 2 | 1 |

| | | | | |
|---|---|---|---|---|
| 2 | Vector of cluster diameters (maximum within cluster distances) | 662.138 | 42.894 | - |
| 3 | Vector of cluster wise within cluster average distances | 184.760 | 42.894 | - |
| 4 | Vector of cluster wise minimum distances of a point in the cluster to a point of another cluster | 46.975 | 46.975 | 49.170 |
| 5 | Vector of cluster wise average distances of a point in the cluster to the points of other clusters | 214.068 | 240.371 | 161.778 |



### 3.3.3. Average linkage method

The Average linkage treats the distances between two clusters as the average distance between all pairs of items where one member of pair belongs to each cluster. The aftereffects of Average linkage technique are spoken to in the Table 3 and the dendrogram for the relating table is as appeared in Figure 5.

Table 3can be utilized to tell that there are 11 germplasm promotions in the first bunch, 29 increases in the second group and 24 increases are there in the third bunch.Characters attributed by the three clusters utilizing Average linkage strategy are recorded in table 6.It tends to be said by looking table that cluster wise minimum distances of a point in the cluster to a point of another cluster are the same for both cluster 1 and cluster 2, with the value 15.013, be that as it may, cluster 3 is having most noteworthy (20.964) for this character. The cluster diameter is estimated to be highest (283.605) for cluster 1.In like manner, cluster wise within cluster average distance was seen as most elevated (116.142) for cluster 1. Cluster wise average distances of a point in the cluster to the points of other clusters is found to be highest for cluster 3 that is 274.149, and cluster 1 was having a nearer value (273.239) to that of cluster 3(Faisal *et al.*2007).

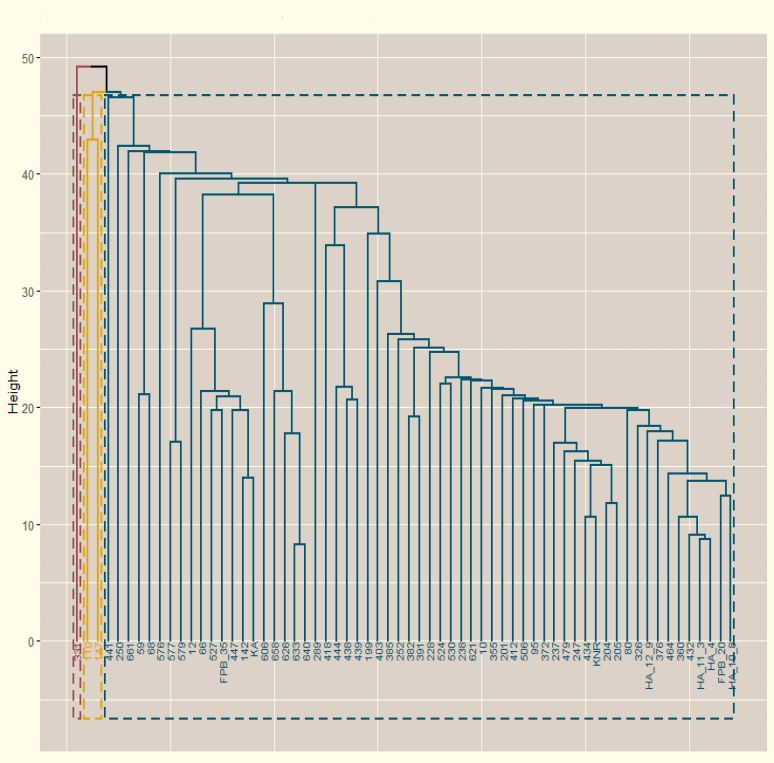


**Fig.3: Dendrogram of Single linkage method for 64 germplasm accessions of Dolichos bean**

**Table 2: Description of the cluster wise characters for Complete linkage method**

| Sl. no. | Characters of individual clusters | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|---|
| 1 | Cluster Size | 11 | 29 | 24 |
| 2 | Vector of cluster diameters (maximum within cluster distances) | 180.743 | 237.261 | 271.404 |
| 3 | Vector of cluster wise within cluster average distances | 78.394 | 80.354 | 120.020 |
| 4 | Vector of cluster wise minimum distances of a point in the cluster to a point of another cluster | 46.975 | 46.975 | 49.170 |
| 5 | Vector of cluster wise average distances of a point in the cluster to the points of other clusters | 214.068 | 240.371 | 161.778 |

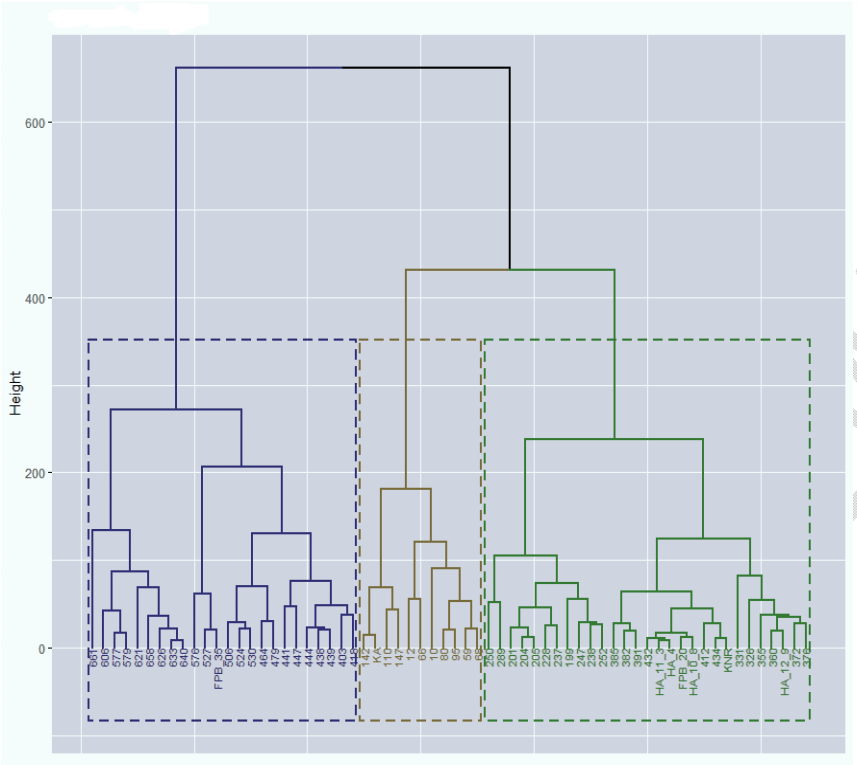


**Fig.4: Dendrogram of complete linkage method for 64 germplasm accessions of Dolichos bean**

**Table 3: Description of the cluster wise characters for Average linkage method**

| Sl.no | Characters of individual clusters | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|---|
| 1 | Cluster Size | 22 | 30 | 12 |
| 2 | Vector of cluster diameters (maximum within cluster distances) | 283.605 | 207.832 | 209.457 |
| 3 | Vector of cluster wise within cluster average distances | 116.142 | 69.869 | 81.244 |
| 4 | Vector of cluster wise minimum distances of a point in the cluster to a point of another cluster | 15.013 | 15.013 | 20.964 |
| 5 | Vector of cluster wise average distances of a point in the cluster to the points of other clusters | 273.235 | 202.712 | 274.149 |

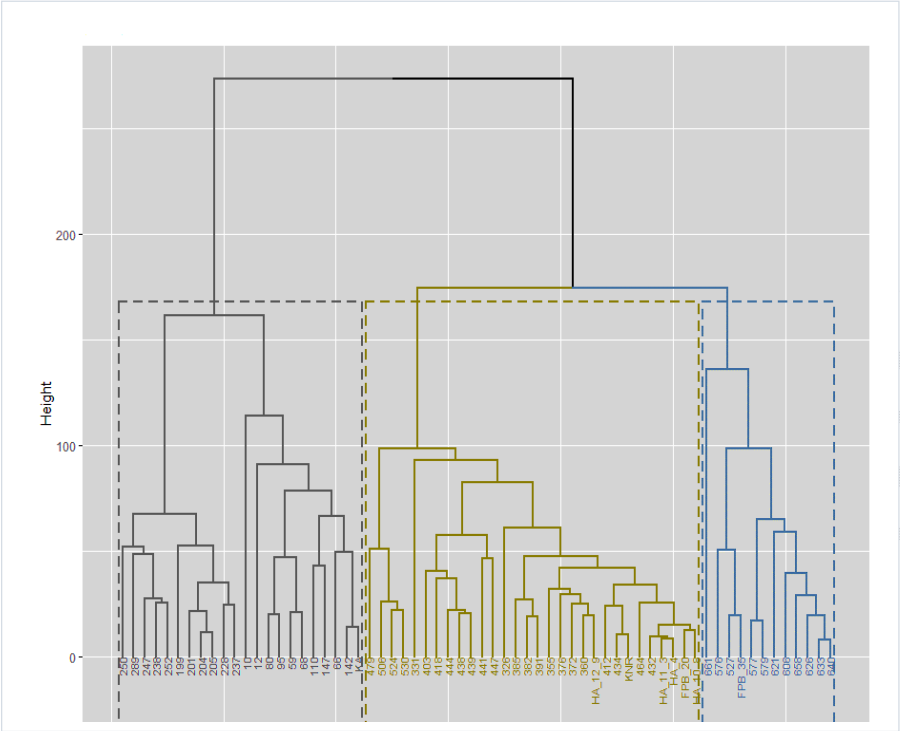


**Fig.5: Dendrogram of Average linkage method for 64 germplasm accessions of Dolichos bean**

**Table 4: Characters of across clusters obtained by Single, Complete and Average linkage method**

| Characters across clusters | Single linkage method | Complete linkage method | Average linkage method |
|---|---|---|---|
| Average distance between clusters | 213.980 | 240.562 | 245.448 |
| Average distance within clusters | 180.256 | 94.892 | 87.908 |
| Number of distances between clusters | 185 | 1279 | 1284 |
| Number of distances within clusters | 1831 | 737 | 732 |
| Maximum cluster diameter | 662.138 | 271.404 | 283.605 |
| Minimum cluster separation | 46.975 | 14.300 | 15.013 |
| Correlation between distances | 0.058 | 0.487 | 0.534 |
| Widest within-cluster gap | 46.531 | 75.577 | 75.577 |
| Dunn index | 0.071 | 0.526 | 0.529 |
| Silhouette Validity Index | -0.227 | 0.410 | 0.461 |

**Table 5: Cluster wise average green seed yield and green pod yield for Average linkage method**

| | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|
| **Average green seed yield (Kg.)** | 69.003 | 60.169 | 82.142 |
| **Average green pod yield (Kg.)** | 118.281 | 110.549 | 152.541 |

From table 4 we can infer that average linkage method of clustering was found to be best compared to other two methods of clustering that is based on cluster validation indicies values *viz.,* dunn index and silhouette validity index was found to be 0.529 and 0.461 respectively. While Table 5 provides an information about average green seed yield and average green pod yield for all the three clusters. Among these clusters, 3$^{rd}$ cluster was found to be having highest value for average green seed yield and green pod yield*i. e.,* 82.142 and 152.541 respectively.

### 3.4. Conclusion

The present study helps in finding the best way to cluster the genotypes of Dolichos bean based on the values of grain yield and its component morphological characters. Besides, it speaks about the influence that the meteorological constraints will have on the grain yield and few other traits. The validation indices used for validating the clusters formed using different methods of clustering have shown that Average linkage method had produced clusters in the best way compared to other clustering methods. The results of validating criterions have shown that Average linkage was having highest value for both the validating indices, forming three clusters with sizes 22, 30 and 12 at a dissimilarity coefficient of 165. Among the three clusters by average linkage method, cluster 3 which had 12 number of accessions was having highest green seed yield (82.142 kg.) and green pod yield (152.541 kg.).

**COMPETING INTERESTS DISCLAIMER:**

Authors have declared that they have no known competing financial interests OR non-financial interests OR personal relationships that could have appeared to influence the work reported in this paper.

### 3.5. Reference


Babu, R.V., Shreya, K., Dangi, K.S., Usharani, G., And Shankar, A.S., 2012, Correlation and path analysis studies in popular rice hybrids of India. *Int. J. Sci. & Res. Pub.,* 2: 1-5.

Dean W. Wichern, Richard A. Johnson, And Richard Johnson, 1988, *Applied Multivariate Statistical Analysis*. Prentice-Hall, New Jersey.

Makeen, K., Abrahim, G., Jan, A. And K. Singh, A.K.,2007, Genetic Variability And correlations studies on yield and its components in mung bean (*Vignaradiata* L. Wilezek). *Agron. J.,***6**(1): 216-218.

Parmar, A.M., Singh, N.P.S., Dhillon And Jamwal, M., 2013, Genetic variability studies for morphological and yield traits in dolichos bean (*Lablab purpureus* L.). *World J. Agric. Sci.,***9**(1): 24-28.

Tiwari Mamta And Misra Bharat, 2011, Application of cluster analysis in agriculture – a review article. *Int. J. Comp. Appl.,* 1: 43 – 47.