International Journal of Agriculture Sciences
ISSN: 0975-3710 & E-ISSN: 0975-9107, Volume 10, Issue 15, 2018, pp.-6801-6805.
Available online at https://www.bioinfopublication.org/jouarchive.php?opt=&jouid=BPJ0000217

Research Article

# MULTIVARIATE CLUSTERING TECHNIQUES: A COMPARISON BASED ON HYBRID TEA GENOTYPES OF ROSE

**ARYA V. CHANDRAN*AND VIJAYARAGHAVA KUMAR**
Department of Agricultural Statistics, College of Agriculture, Vellayani, 695522, Kerala Agricultural University, Thrissur, 680656, Kerala, India
*Corresponding Author: Email - aryavc009@gmail.com

Received: July 23, 2018; Revised: August 10, 2018; Accepted: August 11, 2018; Published: August 15, 2018

**Abstract:** Multivariate clustering technique is an important tool for interpreting the data and to find out natural grouping. Diverse techniques are available there but results are not unique. Study was under taken to compare different clustering techniques. Data on quantitative traits collected from a field experiment on 25 Hybrid Tea genotypes were used for the study. Different hierarchical clustering methods and k-means clustering were compared using measures like Euclidean, Squared Euclidean, Chebychev, City Block and Mahalanobis' $D^2$ statistics. Principal component analysis (PCA) was also carried out and score plot obtained from PCA helps to identify clusters visually. The analysis revealed that clustering obtained from $D^2$ statistics is different from other association measures. Similarity was found among Euclidean and Squared Euclidean distance. Unweighted Pair Group Average Method (UPGMA) and Weighted Pair Group Average Method (WPGMA) gave similar clustering pattern. UPGMA method under Squared Euclidean have minimum SD index.

**Keywords:** Association measures, Clustering methods, PCA

**Citation:** Arya V. Chandran and Vijayaraghava Kumar (2018) Multivariate Clustering Techniques: A Comparison Based on Hybrid Tea Genotypes of Rose. International Journal of Agriculture Sciences, ISSN: 0975-3710 & E-ISSN: 0975-9107, Volume 10, Issue 15, pp.- 6801-6805.


## Introduction

Cluster analysis is multivariate technique which classifies objects based on a set of characteristics in such a way that the resulting clusters have high internal homogeneity and high external heterogeneity [1-2]. Cluster analysis involves measure of similarity, selection of clustering technique and carrying out clustering based on the selected technique. The choice of association measures to be used in cluster analysis has a strong impact on clustering results [3-4]. Several association measures are there that can be used for quantitative data. In this study Euclidean, Squared Euclidean, Chebychev, City Block and Mahalanobis $D^2$ statistics were used as distance measure. The difference between clusters created with Euclidean, Squared Euclidean and Manhattan distance is rather small [5]. Clustering algorithm like single linkage, complete linkage, UPGMA, WPGMA, UPGMC, Ward's method, Tocher method and k- means clustering were compared. Single linkage clustering under different association measures suffering from chaining effect [6] and cluster quality is highly affected by this chaining effect [7]. Complete linkage method and Ward's method gave similar results, while average method is most similar to centroid method [6]. Principal component analysis was also carried out to identify the clusters. SD index were calculated for single linkage, complete linkage and average method.

## Materials and Methods

In this study comparison of different association measures and clustering method were done with the data on quantitative traits of 25 varieties of rose coming under Hybrid Tea group collected Regional Agricultural Research Station (RARS), Ambalavayal, Wayand [Table-1]. The data on quantitative traits such as number of leaves at first flower, number of days to first flower, prickle density (per five cm), flower size (cm), flower weight (g), pedicel length (cm), number of petals flower$^{-1}$ ,size of petals (cm), number of flower plant$^{-1}$/bunch$^{-1}$ were collected for the study.

The genotypes under study need to be test for homogeneity prior to clustering. Multivariate analysis of variance is the method for testing homogeneity of varieties and it involves the technique of analyzing variance and covariance of variables and partioning of these variance into different components [8]. The model for each observation vector is

$$Y = \mu + \alpha + \varepsilon$$

where Y is vector of individual responses, $\mu$ is vector of general mean effect, $\alpha$ is vector of treatment effect and $\varepsilon$ is random error vector which follows $N(0,\Sigma)$.The total dispersion is split up into various components, between genotypes and within genotypes. Wilk's lamda criterion developed through generalized likelihood principle [9] and statistics used for testing the homogeneity of a given set of genotypes is given by V (stat) = -mlog $\Lambda$, where V (stat) is distributed as $\chi 2$ with pq degrees of freedom and $m = n-(p+q+1)/2$, p is number of variables, q is d.f. for variety and n is d.f. for error + variety. Significance of V (stat) shows that the differences between the populations with respect to means of 'p' characters are significant.

The data were subjected to analysis variance corresponding to completely randomized design (CRD) with ANOVA model as

$$X_{ij} = \mu + t_i + e_{ij}, \quad i = 1,2,\ldots,p$$

where $\mu$ is the general mean, $t_i$ is the effect of $i^{th}$ treatment and $e_{ij}$ is the error component with respect to $i^{th}$ character and $e_{ij}$ are normally distributed with mean zero and constant variance. Association measures used for quantitative data are given in [Table-1] [1,10]. Clustering methods includes different linkage methods, Ward's method, Tocher method and k – means clustering. Different linkage methods used for the study are given in [Table-2], where $x \varepsilon A$ and $y \varepsilon B$ Tocher. method and Ward's method are different from linkage methods even though they are hierarchical methods. Ward's method uses analysis of variance approach to calculate the similarity of clusters. Procedure is based on minimizing the loss of information [11]. In Tocher method first cluster is formed by joining objects having smallest distance. A third object having smallest average $D^2$ value from the first two objects is added. The process repeats until an abrupt change in $D^2$ value is noticed [7]. k- means clustering approach is a partitioning approach which reallocate the objects in each iteration [12]. The SD validity index measures the average scattering and total separation of clusters [13]. Scattering is obtained by calculating variance of the clusters and the variance of the complete dataset.

For a compact cluster, variance of the cluster will be smaller than the variance of dataset. Better cluster configuration can be identified by lower SD index.

Table-1 Hybrid Tea genotypes

| Number | Genotypes |
|---|---|
| 1 | Madame George Delbard |
| 2 | Aiswarya |
| 3 | Christ of Colomb |
| 4 | Pink Panther |
| 5 | Roughe Miland |
| 6 | Shrewsbury show |
| 7 | Alaine Souchen |
| 8 | Amara |
| 9 | Fryat |
| 10 | Perfume Perfect |
| 11 | Silver Star |
| 12 | Lincoln Cathedral |
| 13 | A tago |
| 14 | Demestra |
| 15 | Golden Fairy Sport |
| 16 | Mary Jean |
| 17 | Toplesse |
| 18 | Priority Pride |
| 19 | Majestic |
| 20 | Prince Jardiner |
| 21 | Cel b Lau |
| 22 | Lois Wilson |
| 23 | Mom's Rose |
| 24 | Alabama |
| 25 | Josepha |

Table-2 Association measures for quantitative data

| Name | Measure |
|---|---|
| Euclidean | $\sqrt{\sum_{i=1}^{p}(X_i - Y_i)^2}$ |
| Squared Euclidean | $\sum_{i=1}^{p}(X_i - Y_i)^2$ |
| Chebychev | Maximum $\lvert X_i - Y_i \rvert$ |
| City Block | $\sum_{i=1}^{p}\lvert X_i - Y_i \rvert$ |
| Mahalanobis $D^2$ | $(\bar{X}_1 - \bar{X}_2)' W^{-1} (\bar{X}_1 - \bar{X}_2)$ |

Table-3 Different linkage methods

| Methods | Algorithm |
|---|---|
| Single linkage | $d(A, B) = \min d(x, y)$ |
| Complete linkage | $d(A, B) = \max d(x, y)$ |
| UPGMA | $d(A, B) = \frac{\sum_{x \in A}\sum_{y \in B} d(x,y)}{(n_A * n_B)}$ |
| WPGMA | $d(AB), k = \frac{d_{Ak} + d_{Bk}}{2}$ |
| UPGMC | $d(A, B) = d(A_c, B_c)$ |

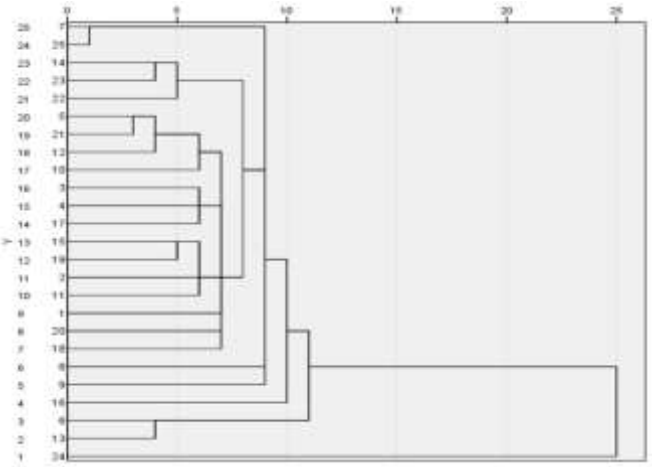

Fig-1 Single linkage – Squared Euclidean

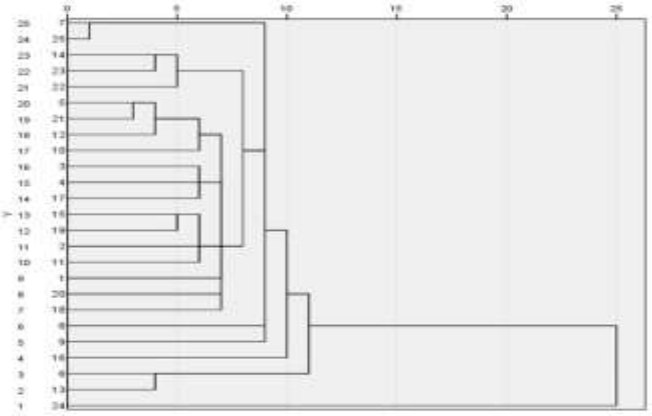

Fig-2 Single linkage – Chebychev distance

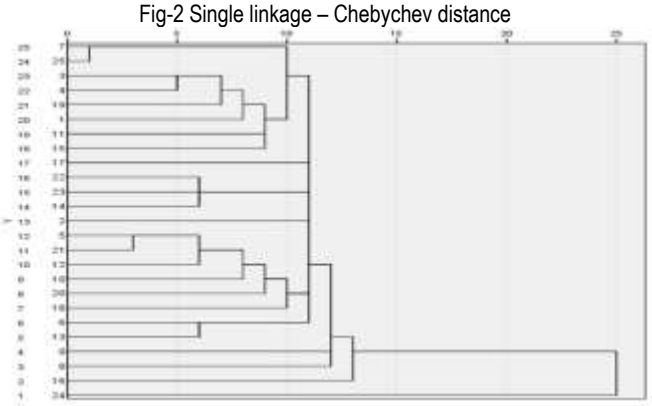

Fig-3 Single linkage – City Block distance

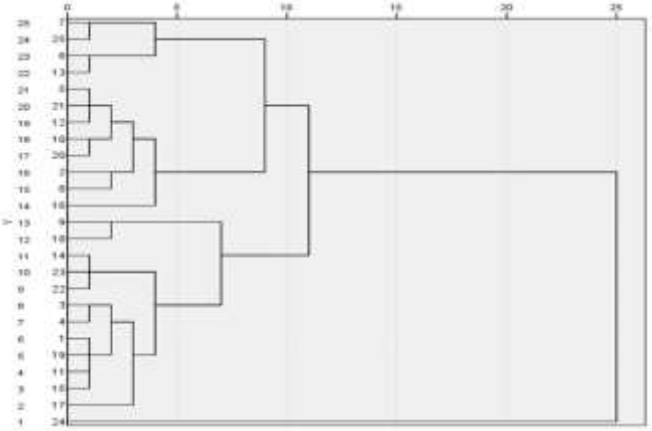

Fig-4 Complete linkage – Squared Euclidean

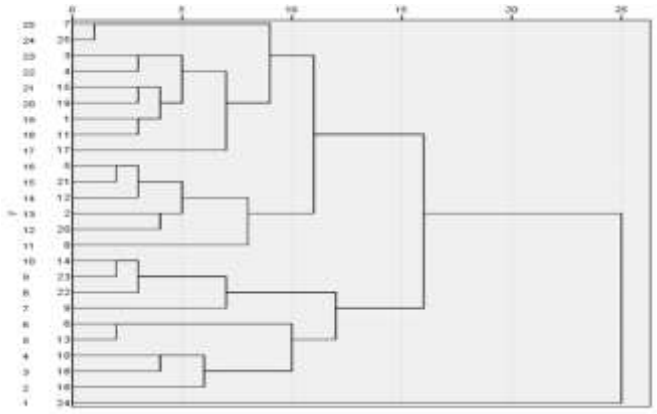

Fig-5 Complete linkage – Chebychev distance

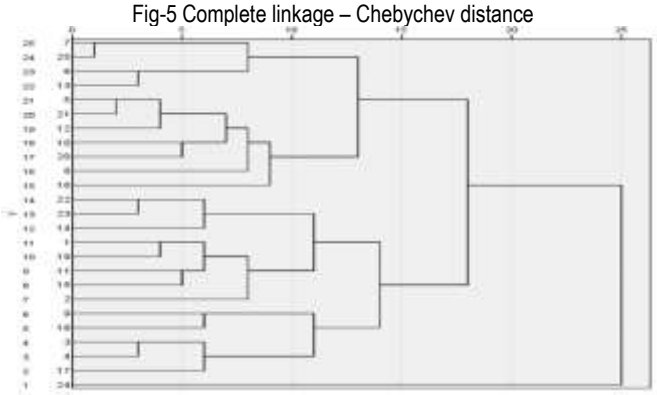

Fig-6 Complete linkage – City Block distance

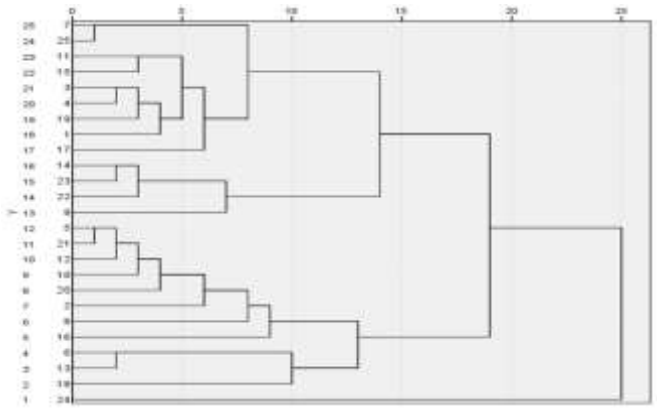

Fig-7 UPGMA – Squared Euclidean

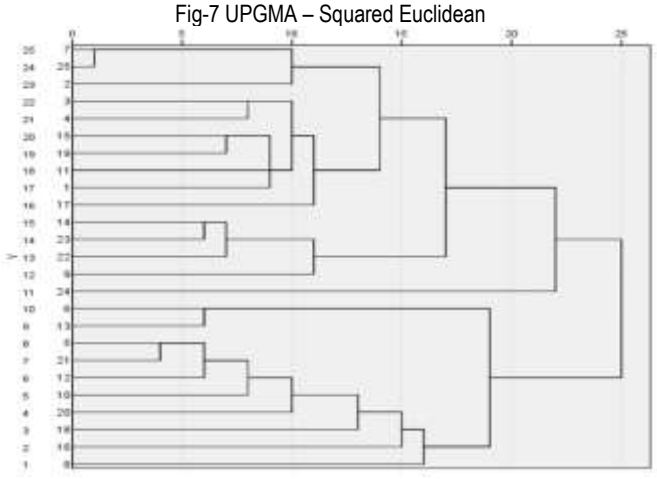

Fig-8 UPGMA – Chebychev distance

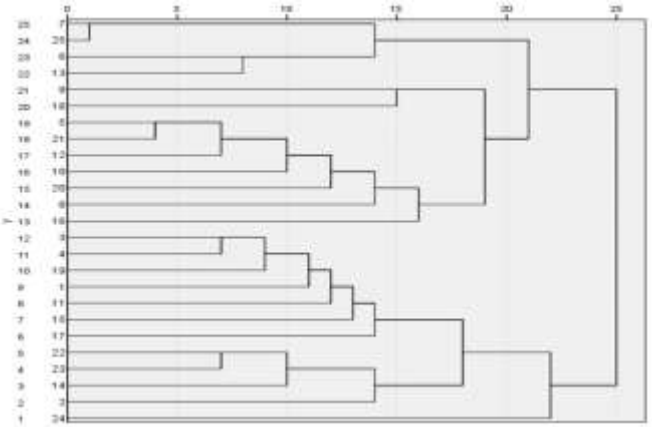

Fig-9 UPGMA – City Block distance

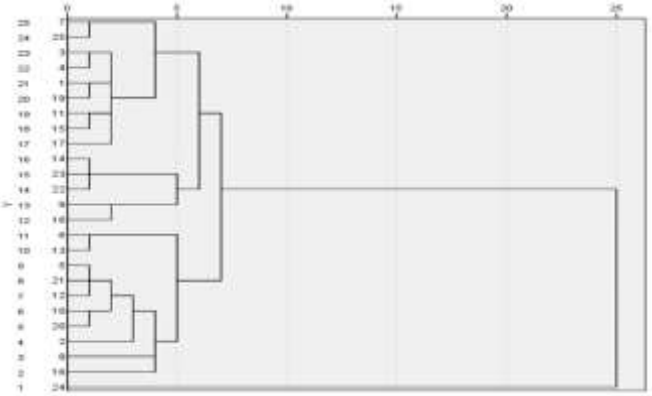

Fig-10 WPGMA – Squared Euclidean

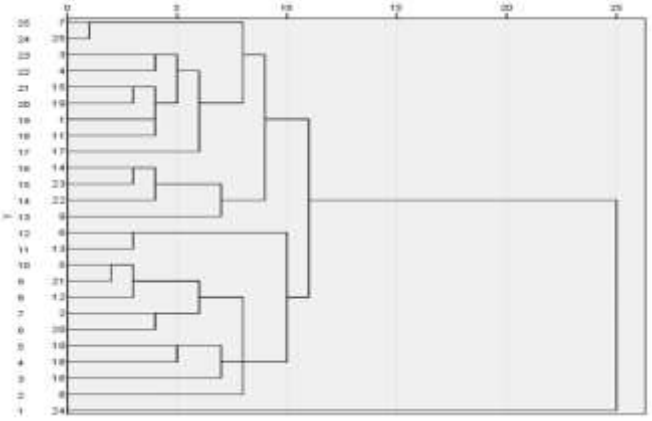

Fig-11 WPGMA – Chebychev distance

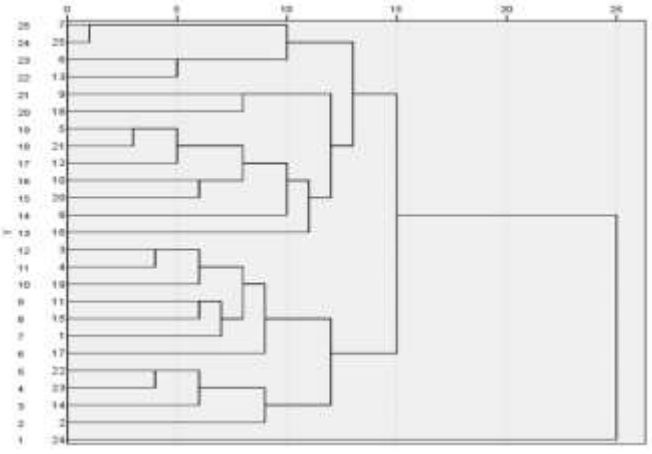

Fig-12 WPGMA – City Block distance

Table-4 Mean values of various characters under study

| Sl. No | Hybrid Tea Genotypes | No. of leaves at first flower | No. of days to first flower | Prickle density/5cm | Flower size (cm) | Flower weight (g) | Pedicel length (cm) | No. of petals/ flower | Size of petals (cm) | No. of flower per plant/bunch |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Madame George Delbard | 25.67 | 195.83 | 6.50 | 36.12 | 3.17 | 4.72 | 18.00 | 5.35 | 1.50 |
| 2 | Aiswarya | 42.50 | 191.67 | 9.33 | 42.38 | 6.87 | 4.47 | 25.33 | 12.92 | 1.33 |
| 3 | Christ of Colomb | 28.50 | 182.50 | 6.00 | 20.14 | 4.98 | 6.92 | 28.67 | 5.70 | 1.33 |
| 4 | Pink Panther | 28.50 | 193.83 | 2.83 | 22.67 | 4.63 | 6.47 | 27.83 | 3.68 | 1.50 |
| 5 | Roughe Miland | 52.67 | 204.67 | 2.33 | 23.97 | 4.51 | 5.84 | 33.00 | 3.40 | 1.50 |
| 6 | Shrewsbury show | 61.67 | 203.17 | 6.50 | 34.63 | 7.62 | 6.67 | 55.00 | 3.66 | 2.00 |
| 7 | Alaine Souchen | 24.33 | 203.17 | 6.33 | 27.46 | 6.09 | 5.85 | 45.17 | 4.37 | 1.67 |
| 8 | Amara | 59.67 | 207.50 | 3.50 | 53.04 | 7.17 | 7.30 | 17.33 | 6.66 | 2.00 |
| 9 | Fryat | 54.17 | 160.00 | 6.33 | 23.81 | 4.89 | 5.85 | 33.67 | 6.79 | 2.00 |
| 10 | Perfume Perfect | 61.50 | 192.33 | 2.00 | 26.41 | 5.00 | 8.07 | 32.50 | 7.41 | 2.00 |
| 11 | Silver Star | 24.67 | 182.83 | 1.83 | 37.86 | 5.60 | 4.92 | 25.50 | 11.60 | 1.50 |
| 12 | Lincoln Cathedral | 57.67 | 211.00 | 3.67 | 23.32 | 4.80 | 4.98 | 23.67 | 3.56 | 1.50 |
| 13 | A tago | 53.33 | 204.67 | 3.17 | 32.65 | 6.99 | 6.25 | 62.33 | 5.15 | 1.67 |
| 14 | Demestra | 46.00 | 164.00 | 4.00 | 41.65 | 1.76 | 4.12 | 18.17 | 1.43 | 2.83 |
| 15 | Golden Fairy Sport | 31.17 | 182.00 | 5.83 | 32.89 | 2.29 | 6.15 | 13.67 | 14.46 | 2.33 |
| 16 | Mary Jean | 82.67 | 197.00 | 11.67 | 32.51 | 6.13 | 5.52 | 23.83 | 7.25 | 2.00 |
| 17 | Toplesse | 17.67 | 170.00 | 2.33 | 22.60 | 5.81 | 5.80 | 22.83 | 9.73 | 1.50 |
| 18 | Priority Pride | 66.17 | 178.17 | 3.33 | 22.26 | 5.19 | 5.72 | 43.50 | 5.40 | 1.33 |
| 19 | Majestic | 26.50 | 190.33 | 4.33 | 23.36 | 7.09 | 5.53 | 14.83 | 6.64 | 2.33 |
| 20 | Prince Jardiner | 55.83 | 191.83 | 3.17 | 33.38 | 3.80 | 5.69 | 18.00 | 8.99 | 2.00 |
| 21 | Cel b Lau | 53.00 | 205.67 | 8.50 | 23.82 | 6.36 | 5.85 | 31.00 | 5.30 | 1.83 |
| 22 | Lois Wilson | 34.33 | 165.83 | 4.33 | 49.08 | 5.98 | 6.48 | 26.00 | 10.41 | 1.00 |
| 23 | Mom's Rose | 43.67 | 166.50 | 4.00 | 44.00 | 5.06 | 4.93 | 23.00 | 8.97 | 1.33 |
| 24 | Alabama | 65.67 | 110.00 | 3.83 | 42.32 | 5.83 | 5.98 | 14.67 | 17.69 | 1.33 |
| 25 | Josepha | 25.33 | 203.17 | 6.33 | 26.61 | 6.06 | 6.02 | 45.17 | 4.43 | 1.50 |

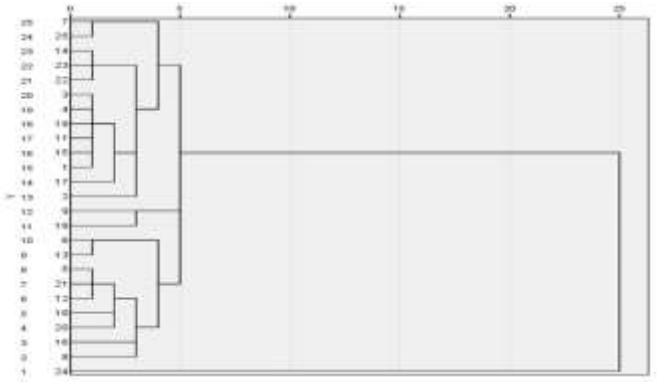

Fig-13 UPGMC – Squared Euclidean

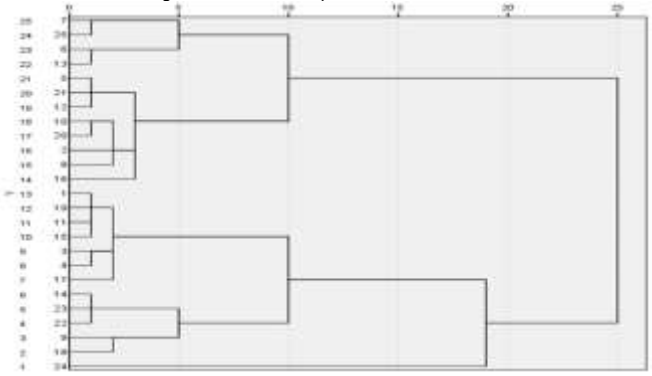

Fig-14 Ward's method - Squared Euclidean

Table-5 Clustering under $D^2$ statistics

| Cluster | Name of the Genotypes |
|---|---|
| I | 7, 25, 5, 21, 12, 4 |
| II | 22, 23, 11, 15, 2 |
| III | 10, 18, 20, 16 |
| IV | 1, 19, 3, 17 |
| V | 6, 13 |
| VI | 9, 14 |
| VII | 8 |
| VIII | 24 |

Table-6 Percentage contribution of characters towards divergence

| Character | Percentage contribution to variance |
|---|---|
| No. of leaves at first flower | 19.00 |
| No. of days to first flower | 40.67 |
| Prickle density/5cm | 0.33 |
| Flower size (cm) | 12.00 |
| Flower weight (g) | 0.00 |
| Pedicel length (cm) | 0.00 |
| No. of petals/ flower | 16.33 |
| Size of petals (cm) | 11.67 |
| No. of flower per plant/bunch | 0.00 |

Table-7 Clustering under k-means clustering

| Cluster no. | Members |
|---|---|
| 1 | 1,3, 11, 15, 17, 19 |
| 2 | 5, 10, 12, 20, 21 |
| 3 | 14, 22, 23 |
| 4 | 4, 7, 25 |
| 5 | 9, 18 |
| 6 | 6, 13 |
| 7 | 2,8 |
| 8 | 16 |
| 9 | 24 |

Table-8 PCA of Hybrid Tea genotypes

| Principal component | Percentage variance | Cumulative variance |
|---|---|---|
| Component 1 | 50.19 | 50.19 |
| Component 2 | 29.012 | 79.21 |
| Component 3 | 12.71 | 91.91 |

Table-9 Principal component loading of different characters

| Characters | PC 1 | PC 2 | PC 3 |
|---|---|---|---|
| No. of leaves at first flower | -0.028 | 0.983 | -0.123 |
| No. of days to first flower | 0.940 | -0.011 | -0.328 |
| Prickle density/5cm | 0.017 | 0.026 | -0.024 |
| Flower size (cm) | -0.146 | 0.076 | -0.335 |
| Flower weight (g) | 0.016 | 0.017 | 0.030 |
| Pedicel length (cm) | 0.006 | 0.009 | 0.015 |
| No. of petals/ flower | 0.286 | 0.163 | 0.872 |
| Size of petals (cm) | -0.110 | -0.013 | -0.059 |
| No. of flower per plant/bunch | 0.001 | 0.003 | -0.010 |

## Results

Analysis of variance was done for each of the character under study which shows significant difference among different genotypes with respect to character. The mean values of various characters corresponding to different genotype are shown in the [Table-4]. The total variation was split up into variation due to between groups and within groups by analysis of dispersion method. The Wilk's lamda value obtained was 0.004 which was found to be significant. The results shows that the difference between the varietal means with respect to character under study. Dendrogram obtained from different clustering methods under different association measures are given in [Fig-1] to [Fig-14]. The result of different clustering techniques based on Squared Euclidean results gave approximately same result as that of Euclidean distance. So the result corresponding to Euclidean distance is not presented. UPGMC and Ward's method were performed only using the Squared Euclidean measure as this method give valid result only for that distance measure. Clustering membership of different genotypes under $D^2$ statistics is given in the [Table-5]. Contribution of characters towards total divergence obtained from $D^2$ analysis is given in the [Table-6]. k – means clustering grouped the genotypes into 9 clusters. Clustering of genotypes under 9 clusters are given in [Table-7]. Principal component analysis was carried out with nine characters [Table-8]. Table of component loading shows the importance characters towards variance [Table-9]. [Fig-15] shows the three dimensional score plot obtained from the PCA which identifies clusters visually.

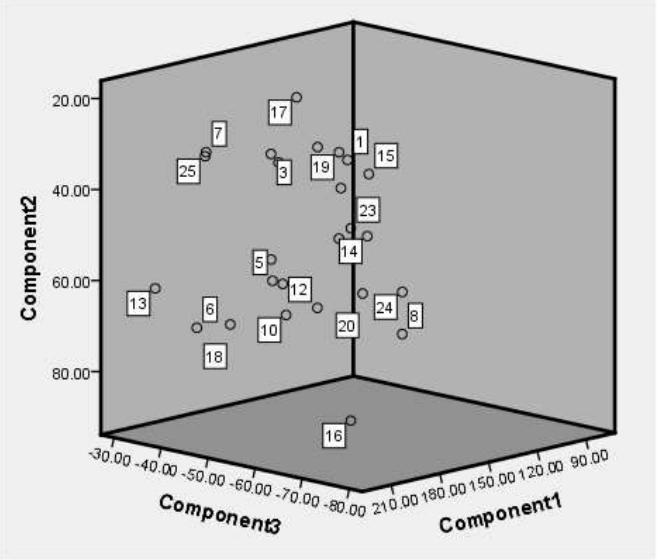


Fig-15 Three dimensional score plot

## Discussion

Analysis of variance (ANOVA) using all quantitative characters under study revealed significant difference among different genotypes with respect to each character. Multivariate analysis of variance (MANOVA) revealed difference among the cultivar means respect to the characters. Single linkage clustering under different distance measures tends to create a set of one or two clusters including majority of the genotypes and the remaining are singletons. Single linkage clustering tends to produce long chain types clusters as opposed to bunched clusters and the single linkage algorithm suffers chaining effect. Among other clustering algorithms, complete linkage method and Ward's clustering method showed similar results under Squared Euclidean distance. UPGMA, WPGMA and UPGMC methods under Squared Euclidean method gave comparable results. Clustering using UPGMA and WPGMA method gives almost same clustering pattern under different distance. Results obtained from k means clustering are comparable with results obtained from hierarchical clustering except for Single linkage clustering. A certain degree of similarity was observed between k means and $D^2$ analysis but not to up that between other clustering methods. Comparison among single linkage, complete linkage and Average linkage under different association measures using SD index revealed that Average linkage method under Squared Euclidean was best with SD index 0.651. Clustering pattern observed from score plot of PCA is comparable with the pattern obtained from quantitative data especially with $D^2$ analysis. PCA indicated that the characters number of days to first flower, number of leaves at first flower and number of petals/ flower have highest contribution to the variance. It is similar to the result obtained from $D^2$ analysis.

## Conclusion

Clustering obtained from different association measures and clustering methods are different. It is possible to compare different measures and can exclude inappropriate methods. Single linkage under different distance measures suffering from chaining effect. Among single linkage, complete linkage and average linkage methods average linkage under Squared Euclidean distance was found to be best for quantitative data.

**Application of research:** Research helps to identify appropriate clustering method for quantitative data.
**Research Category**: Multivariate analysis, Cluster analysis

**Abbreviations**: cm: centimeter, g: gram, UPGMA: Unweighted Pair Group Average Method, WPGMA: Weighted Pair Group Average Method

**Acknowledgement / Funding**: Author thankful to College of Agriculture, Vellayani, 695522, Kerala Agricultural University, Thrissur, 680656, Kerala, India

***Research Guide or Chairperson of research: Dr Vijayaraghava Kumar**
University: Kerala Agricultural University, Thrissur, 680656, Kerala, India
Research project name or number: MSc Thesis

**Author Contributions**: All author equally contributed
**Author statement**: All authors read, reviewed, agree and approved the final manuscript

**Conflict of Interest**: None declared

**Ethical approval**: This article does not contain any studies with human participants or animals performed by any of the authors.

## References

[1] Hair J.F., Black W.C., Babin B.J., Anderson R.E. (2015) *Multivariate Data Analysis (7th Ed.), Pearson Education Ltd, New York,*415-474.
[2] Ghuman, S.S. (2016) *International Journal of Computer Science and Mobile Computing*, 5, 524-530.
[3] Duarte M.C., Santos J.B., Melo, L.C. (1999) *Genetics and Molecular Biology,* 22, 427-432.
[4] Jackson A.A., Somers K.M., Harvey H.H. (1989) *The American Naturalist*, 133, 436-453.
[5] Dahal S. (2015) *Effect of Different Distance Measures in Result of Cluster Analysis, M.Sc. thesis, Aalto University School of Engineering, Finland*, 77.
[6] Kuiper F. K., Fisher, L.A. (1975) *Biometrics*,31, 777-783.
[7] Oyang Y., Chen C.Y., Yang T.W. (2001) *Principles of Data Mining and Knowledge Discovery. PKDD 2001. Lecture Notes in Computer Science, vol 2168. Springer, Berlin, Heidelberg*
[8] Rao C.R. (1952) *Advanced Statistical Methods in Biometric research, John Wiley and Sons, Inc. New York*, 246-250.
[9] Wilks S.S. (1932) *Biometrika*, 24, 471.
[10] Krzanowski W., Marriott F. (1995) *Kendall's Library of Statistics 2 Multivariate analysis, John Wiley and Sons, Inc. New York,* 61-94.
[11] Ward J. H. (1963) *J. of American Statistical Association*, 58, 69-78.
[12] MacQueen J.B. (1967) *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press*, 1,281-297.
[13] Halkidi M., Vazirgiannis M., Batistakis Y. (2000) *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery,* 265- 276.