Original Research Article

Cluster analysis approach for studying performance of hybrids in varietal experiments

P Sai Shankar[1*], J V Narasimham[1], Ananthan Gopal[1], Dr. B. S Kulkarni[2]

[1]Reliance Technology Group, Reliance Industries Limited, Hyderabad, Telangana, INDIA.
[2]Consultant, Reliance Technology Group, Reliance Industries Limited, Hyderabad, Telangana, INDIA.
Email: Sai.Pothakani@ril.com

Abstract

In the context of analysis of field experimental data, an attempt has been made to explore the feasibility of Cluster Analysis approach in grouping of treatment means. Cluster analysis approach has an edge over the conventional ANOVA based groupings in providing non-overlapping groups of treatments that enables a researcher to identify the treatments with similar performance. The application has to be developed such that it provides clusters (groups) of treatments that are distinctly different from one another but homogeneous within themselves. A procedure based on ANOVA has been proposed that provide clustering with these characteristics. The approach was applied for obtaining groupings of 16 Non Edible Oil seed hybrids on the basis of crop yield recorded from the varietal trial conducted during 2013-14 at Andhra Pradesh, India. The experiment was conducted in Split plot design with 3 replications and 16 Entries (14 Hybrids and 2 Open Pollinated Varieties as Controls) with 3 levels of Irrigation (Full Irrigation, Supportive Irrigation and Rainfed). ANOVA based analysis of experimental data revealed significant response of only the hybrids. The comparison of hybrids on the basis of the conventional Least significance difference(LSD) exhibited overlapping groups. Application of the proposed the cluster analysis approach has resulted in 7 distinct groups (Clusters) of hybrids. The cluster with single hybrid S0011 was identified as relatively the best performing hybrid.
Key Words: ANOVA, Cluster Analysis, Average Linkage Method.

[*]**Address for Correspondence:**
P Sai Shankar, Reliance Technology Group, Reliance Industries Limited, Hyderabad, Telangana, INDIA.
Email: Sai.Pothakani@ril.com
Received Date: 03/10/2017 Revised Date: 18/11/2017 Accepted Date: 09/12/2017

Access this article online

Quick Response Code:

Website: www.statperson.com

Accessed Date: 10 January 2018

INTRODUCTION

Analysis of experimental data from an experimental design involves a two-step procedure:

1. Obtaining Analysis of Variance (ANOVA) of the design that draws an overall inference about the significant differences of the treatments through F-test and then
2. Comparing pair-wise treatment means to isolate groups of treatments that are significantly different from one another. This is done by applying the well-known multiple comparison tests such as Fisher's Least Significant Difference (LSD) or Tukey's procedures, provided the F-test for treatments in ANOVA reveals significant differences.

In certain situations, these comparisons may lead to overlapping groupings, thus it may not be possible to clearly distinguish between the performances of the treatments (McLachlan, 2001). Cluster Analysis approach has been advocated as an alternative to multiple comparison test procedures to deal with the situation of obtaining non-overlapping groups of treatments (*See for instance*- Scott and Knott, 1974; Willavise*et al*, 1980; Madden *et al*, 1982; Basford and McLachlan, 1985; Jolliffe *et al*, 1989; Bautista *et al*, 1997). The approach has a scientific base of classifying the similar objects (treatments in present situation) on the basis of the relative distances. Cluster analysis approach is also commonly applied in plant breeding trials to study the

How to site this article:P Sai Shankar, J V Narasimham, Ananthan Gopal, B S Kulkarni. Cluster analysis approach for studying performance of hybrids in varietal experiments. *International Journal of Statistika and Mathemtika*. November to January 2018; 25(1): 12-15. http://www.statperson.com

genetic divergence of genotypes by the Tocher's Method (Singh and Choudhary, 1996). *For a detailed exposure on the various clustering methods, see for instance* Mardia *et al* (1995) and Everitt*et al* (2011). The only limitation of the approach is that there is no rational procedure that decides the optimality in the number of clusters that possesses the characteristics of distinctness *between* the clusters and homogeneity (as far as possible) *within* the clusters. This may be achieved by manual iterations to find out the optimum number of clusters, thus making the procedure a little cumbersome. In the present study, an attempt has been made to explore the suitability of cluster analysis approach with emphasis on optimizing the cluster formation that possesses the above characteristics. The approach was applied in the context of evaluating the yield performance of Non Edible Oil seed hybridsin the varietal trial conducted at Andhra Pradesh, India.

MATERIAL AND METHODS

The Approach: In the context of field experimental data, the researcher's expectation is that the analysis of data may provide non-overlapping groups of treatment means that will facilitate in distinguishing the relative performance of treatments. The conventional ANOVA based analysis on the basis of multiple comparison test procedures many at times provide overlapping groups of treatments (Bautista *et al*, 1997). As an alternative, an approach based on cluster analysis has been proposed.

Cluster Analysis Approach: Suppose that the replication-wise data on *k*-varieties recorded from an experiment (design) is summarized in the form of varietal means. These *k*-varietal means corresponding to either a single variable or multiple variables can be then subjected to cluster analysis. Cluster analysis is a well-known multivariate procedure for classifying the objects on the basis of observations recorded either on single or multiple characters into distinct groups, which are referred as Clusters. The classification is on the basis of relative distances of the objects, *i.e.,* varieties in the present situation, through a step-wise procedure (*algorithm)*. Among the various methods of clustering, the Average Linkage method is commonly preferred due to the less subjectivity involved in the cluster formation (Kulkarni and Damodar Reddy, 1994). The outcome of the analysis is a tree-like diagram, known as *Dendrogram*, which exhibits the connectivity of the objects on the basis of the relative distances. The diagram helps in identifying the clusters (*The details of procedures is avoided, as these are well known and accepted)*. The only limitation of the approach is that the number of clusters in which the objects are to be classified, cannot be pre-decided. The intention of applying the procedure to the experimental data is to obtain non-overlapping and distinct groups of varieties, which are different from one another but homogeneous (as far as possible) *within* the groups. Hence the following ANOVA based procedure has been proposed to rationalize the cluster formation: Cluster analysis provides clusters (groups) of objects that are distinct. *Statistically,* it implies that there is significant difference between the clusters summarized on the basis of cluster means. It is likely that the number of objects under each cluster may vary (*i.e.,* unequal). Hence, the distinctiveness of the clusters can be verified on the basis of cluster mean values, by applying ANOVA with One-Way Classification. Here, the *Treatments* to be compared are the Clusters and the observations are the responses (such as crop yield) of varieties classified in the cluster. In the present context, these are the Seed Weight/Plant (grams) recorded corresponding to the varieties. It is possible that the cluster means may exhibit statistically significant differences when the treatments are classified in small number of clusters. This, however, will not ensure the homogeneity within the cluster. Hence, the ANOVA procedure for analyzing cluster mean values has to be repeated with several choices for number of clusters, manually, till any pair of the clusters means arranged in descending order of magnitude exhibit statistically non-significant differences.

Experiment and Data: The study is based on the experimental yield data corresponding to 16 Non Edible Oil seed entries involved in the experimental trial conducted at Andhra Pradesh. The experiment was conducted during 2013-14 in Split Plot Design to study the performance of the Hybrids under 3 Irrigation Treatments. The details of the experiment are:

Main Plot Treatments: 3

T-1: Full Irrigation; T-2: Supportive Irrigation and T-3: Rain-fed.

Sub Plot Treatments: 16 Entries (14 Hybrids and 2 Open Pollinated Varieties)

Replications: 3.

The performance of hybrids was evaluated on the basis of Crop Yield (Seed Weight / Plant (gms.)). The analysis of data was carried out by using MINITAB (*Version: 16)*.The Average Linkage method of clustering was applied to the experimental yield data of 16 Non Edible Oil seed hybrids. The analysis was also carried out with conventional ANOVA based approach, for comparison with the proposed approach.

RESULTS

The analysis of Split Plot experimental data on crop yield (Seed Weight / Plant (gms.)) is presented in Table-1:

Table 1: Analysis of Variance (ANOVA) - Seed Weight / Plant (gms.)

Source	Df	Mean Square	F-Statistic
Replications	2	12421	0.65
Irrigation Treatments	2	10095	0.52
Error (a)	4	19251	
Hybrids	15	9779	6.76**
Irrigation * Hybrids	30	2006	1.39
Error (b)	90	1447	

*(** Significant at 0.05 level of probability)*

The ANOVA reveals that the response of Hybrids alone is statistically significant; while the response of Irrigation Treatments as well as the Interaction of Irrigation and Hybrids was statistically non-significant. The analysis thus led to pairwise comparison of hybrids in the basis of average (mean) yields by applying Fisher's Least significant difference (LSD) procedure. The results of this conventional procedure of comparison of hybrids are presented in Table- 3. The Average yield data of hybrids was also subjected to the Cluster analysis approach as outlined above. The approach was applied with 6 choices for cluster formation. These were 3, 4, 5, 6, 7 and 8 clusters. The One-way ANOVA of cluster data on hybrid yield indicated that cluster means of hybrids, in general, exhibited significant differences in case of small number of clusters such as 3 or 4 clusters. However, these clusters, though significantly different, were not homogeneous, *within* themselves. Specifically, it was observed that the clusters means exhibited *between-cluster significance* and *within-cluster homogeneity* when the clustering was done with 7 clusters; while beyond 7 clusters, certain cluster means exhibited non-significant differences The analysis thus indicated that the *distinct (and homogeneous) clusters* can be formed by classifying the yield data of hybrids in 7 clusters.The ANOVA based results that led to distinctness of the clusters are presented in Table-2.Specifically, the results are presented corresponding to classification of hybrids into only the final 7 and 8 clusters.

Table 2: Cluster Analysis approach / Performance of clusters

Classification with 7 Clusters		Classification with 8 Clusters	
Cluster No	Cluster Means	Cluster No	Cluster Means
C1	141.14 a	C1	141.14 a
C2	108.31 b	C2	108.31 b
C3	85.19 c	C3	90.37 c
C4	63.91 d	C4	80.00 c
C5	48.36 e	C5	63.91 d
C6	36.74 f	C6	48.36 e
C7	20.95 g	C7	36.74 f
		C8	20.95 g

(Mean Values with different letters indicate significant at p = 0.05)

It can be observed from Table-2 that the classification based on 7 clusters resulted in distinct (non-overlapping) clusters; while the one based 8 clusters involved a pair of cluster means (C3, C4) which is statistically non significant. This analysis indicated that the hybrids can be classified into 7 clusters that possesses the characteristics of distinctness (non overlapping) and homogeneity within the clusters. Hence the comparison of hybrids was carried out on the basis of 7 cluster classification and the results are presented in Table- 4.

Table 3: Comparison of Hybrids: Conventional Approach

HYBRIDS	Seed Wt/plant	Grouping
S 0011	141.14	A
S 0014	112.46	A B
S 0005	104.16	B C
S 0009	90.37	B C D
S 0007	80.00	B C D E
S 0010	68.52	B C D E
S 0001	64.01	B C D E
S 0003	63.08	B C D E
S 0004	60.02	B C D E
S 0008	51.27	B C D E
S 0012	50.26	B C D E
S 0002	46.19	C D E
S 0013	45.70	C D E
S 0006	36.74	D E
R 0001	23.05	E
R 0078	18.84	E
LSD (Var)	**35.62**	

Table 4: Comparison of Hybrids: Cluster Analysis Approach

Cluster	Hybrid	Seed Wt/plant (gms)	Cluster Mean
1	S 0011	141.14	141.14
2	S 0014	112.46	108.31
	S 0005	104.16	
3	S 0009	90.37	85.19
	S 0007	80.00	
	S 0010	68.52	
4	S 0001	64.01	63.91
	S 0003	63.08	
	S 0004	60.02	
5	S 0008	51.27	48.36
	S 0012	50.26	
	S 0002	46.19	
	S 0013	45.70	
6	S 0006	36.74	29.90
	R 0001	23.05	
7	R 0078	18.84	18.84

It can be observed from Tables-3 that the conventional approach based on LSD has led to several overlapping combinations of hybrids. Specifically in the context of hybrids with relatively higher mean yields, the pair S0011 and S0014 were on par (Statistically non-significant). However, S0014 of this group exhibited non-significant difference with S0005, when S0011 was significantly different from S0005.In contrast, the classification based on Cluster analysis approach has resulted in non-overlapping of hybrids (Table-4). Specifically, the hybrid with relatively highest average yield i.e., S0011 was alone classified in Cluster 1. Similarly, R0078 with relatively

lowest average yield was classified in Cluster 7. Cluster 2, Cluster 3 and Cluster 6 were formed with three hybrids each; whereas Cluster 4 and Cluster 5 were formed with four hybrids. These distinct groups of hybrids i.e., the seven clusters indicate that the hybrids within these clusters are significantly different from those in other Clusters. The final outcome of the analysis is that S0011 can be identified as the sole relatively best performing Hybrid.

CONCLUSIONS

1. In the context of analysis of yield data on Non Edible Oil seed varietal trial the cluster analysis approach was found to be an effective approach for obtaining non-overlapping groups of hybrids.
2. The conventional approach based on LSD was found to give several overlapping groups of Hybrids.
3. Performance of hybrids was distinctly summarized in seven clusters that enabled us to identify hybrids with similar yield levels.
4. The hybrid S0011 was distinctly identified as the best performing hybrid.

REFERENCES

1. Basford, K.E. and McLachlan, G.J. (1985) Cluster Analysis in Randomized Complete Block Design, Communications in Statistics- Theory and Methods, 14: 451-463
2. Bautista, M.G., Smith, D.W. and Steiner, R.L. (1997) A Cluster -Based Approach to Means Separation, Journal of Agricultural, Biological and Environmental Statistics, 2:179-197
3. Everitt, B.S., Landau, S., Leese, M. and Stahl, D. (2011) Cluster Analysis, John Wiley and Sons Ltd., West Sussex, PO1985Q, UK
4. Jolliffe, I.T., Allen, O.B. and Christie, B.R. (1989) Comparison of Variety Means using Cluster Analysis and Dendrograms, Experimental Agriculture, 25: 259-269
5. Kulkarni, B.S. and Damodar Reddy, D. (1994) Cluster Analysis Approach for Classification of Andhra Pradesh on the basis of Rainfall, MAUSAM, 45: 325-332
6. Madden, L.V., Knoke, J.K. and Louie, R. (1982) Considerations for the use of Multiple Comparison Procedures in Phytopathological Investigations, Phytopathology, 72: 1015-1017
7. Mardia, K.V., Kent, J.T. and Bibby, J.M. (1995) Multivariate Analysis, Academic Press, Waltham
8. McLachlan, G.J. (2001) Letter to Editor, Journal of Agricultural, Biological and Environmental Statistics, 6: 302-304
9. Scott, A. J. and Knott, M. (1974)A Cluster Analysis Method for Grouping Means in the Analysis of Variance, Biometrics, 30: 507–512
10. Singh, R.K. and Choudhary, B.D. (1996) Biometrical Methods in Quantitative Genetic Analysis, Kalyani Publishers, New Delhi
11. Willavise, S.A., Carmer, S.G. and Walker, W.M. (1980) Evaluation of Cluster Analysis for Comparing Treatment Means, Agronomy Journal, 72: 317-320

Source of Support: None Declared
Conflict of Interest: None Declared