Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 65(3) 2011 339-346

Statistical Modeling to Group Villages Based on Soil Parameters

A. Rajarathinam[1], A.N. Khokhar[2], P.R. Vaishnav[2] and S.K. Dixit[2]
[1]*Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu*
[2]*Anand Agricultural University, Anand, Gujarat*

Received 14 October 2009; Revised 02 August 2011; Accepted 02 August 2011

SUMMARY

An empirical investigation was carried out to study the variation in five soil parameters, including pH, electrical conductivity (EC), organic carbon (OC), available phosphorus (P) and potassium (K). All of these are routinely analysed in soil-testing laboratories across the country to test farm soil from various villages and to group these villages according to patterns in these parameters. The data on the five soil parameters pertaining to 47 villages of Godhara taluka in the state of Gujarat were obtained from the Soil Health Card (SHC) scheme. These soil parameters were subjected to various statistical analyses. An analysis of variance (ANOVA) showed that variations in pH, EC, OC, P and K among the villages were highly significant; that is, these individual parameters were significantly different across the villages. A multivariate analysis of variance (MANOVA) test revealed a significant variation between the villages when all the five soil parameters were considered simultaneously. Though all the soil parameters were found to be significant both individually (ANOVA) and together (MANOVA), the clustered variation was largely due to variations in OC, EC and P as confirmed by Ward's method as well as k-means clustering. Three clusters were identified such that there was homogeneity within the clusters and heterogeneity between the clusters. These groupings can be used to prepare fertility maps and to develop effective soil improvement programmes. Though distinct clusters could be identified in Godhara taluka, geographical closeness among the villages in a given cluster was not evident, indicating that the factors influencing the soil parameters are geographically well dispersed.

Keywords : ANOVA, MANOVA, Pivotal condensation, Squared Euclidean distance, Ward's Method, k-means clustering.

1. INTRODUCTION

A general problem that faces many researchers involves the organization of large amounts of observed data into meaningful structures. The organisation of data assists researchers in implementing appropriate sampling techniques and, therefore, ensures the reliability of data in accordance with the objectives of the research programme. It also helps in reducing costs associated with data collection and it increases the efficiency of the entire research programme. If organized data are statistically treated using appropriate methods, meaningful and pragmatic conclusions can be derived. The soil health data bank which was recently developed by the Gujarat State Agricultural Department as part of the Soil Health Card (SHC) programme is one such voluminous set of data. By considering variability in these data, various suitable statistical techniques can be used to derive possible conclusions.

Keeping in mind the importance of analyzing variability patterns in soil data, we attempt to group villages based on five soil parameters, including pH, electrical conductivity (EC), organic carbon (C), phosphorus (P) and potassium (K). These are the parameters most commonly analysed by soil-testing laboratories across the country. An analysis of OC as well as available P and K help to determine the

*Corresponding author : A. Rajarathinam
E-mail address : arrathinam@yahoo.com

recommended rate of N, P and K fertilizers. EC and pH provide guidance as to the choice of crops and the soil management practices that can be used to enhance soil productivity.

The specific objectives of the study are as follows.

(i) To test the significance of variation in each soil parameter between the villages in Godhara taluka using Analysis of variance (ANOVA).

(ii) To study the variability pattern of all the five soil parameters considered simultaneously between the villages of Godhara taluka using Multivariate analysis of variance (MANOVA).

(iii) To group villages into clusters based on the variability patterns of the soil parameters using Ward's method and k-means clustering and to study the characteristics of these clusters.

2. MATERIALS AND METHODS

An attempt was made to group the 47 villages of Godhara taluka based on five soil parameters namely EC (ds/m); OC (%); P (P_2O_5 kg/ha); K (K_2O in kg/ha), using Ward's method and k-means clustering method. In each village, 20 soil samples are collected.

Analysis of Variance

An ANOVA technique (Rao 1952) is employed to test the significance of the variation in each parameter between the villages. We use the following model (Rao 1952)

$$y_{ij} = \mu + v_i + e_{ij} \qquad (2.1)$$
$$i = 1, 2, 3, \ldots, 47;\ j = 1, 2, 3, \ldots, 20$$

where y_{ij} is the status of the soil parameter in the j^{th} village of the i^{th} soil sample, μ is average status, v_i is the status in the i^{th} village and e_{ij}'s are random error which follow a normal distribution with mean zero and constant variance σ^2.

Multivariate Analysis of Variance

To test the significance of variation among all the five parameters considered simultaneously a MANOVA technique (Johnson and Wichern 2002) is employed. The MANOVA model for comparing the population mean vectors (g = 47) is as follows.

$$Y_{ij} = \mu + V_i + E_{ij} \qquad (2.2)$$
$$i = 1, 2, 3, \ldots, 47;\ j = 1, 2, 3, \ldots, 20$$

where E_{ij} is vector of random errors distributed as $N_p(0, \Sigma)$ (p = 1, 2, 3, 4, 5). Here the parameter vector μ is the overall mean and V_i represents the status of the soil parameters in the i^{th} group. According to the model in (2.2), each component of the observation vector Y_{ij} satisfies the univariate model (2.1) and the variance covariance matrix Σ is same for all populations.

Cluster Analysis

In order to address within-variability of the village mean values, different soil parameter mean values are converted into uncorrelated variables using the pivotal condensation method (Rao 1952). The transformed uncorrelated variables are used to group the villages with the Ward's method which involves squared Euclidean distance and the k-means clustering method. The optimum numbers of clusters is calculated based on cluster selection criteria enumerated in Aldenderfer and Blashfield (1984).

Inter- and Intra-Cluster Distances

After the formation of the clusters, inter- and intra-cluster D^2 values are calculated, using averaged individual D^2 values. The square root of these D^2 is used to indicate inter- and intra-cluster distances. The cluster means for all characters are computed using the character means for the villages included in the clusters.

Estimation of intra- and inter-cluster variance for different characters

An un-weighted analysis of variance using the mean value of different characters is implemented (Dixit 1984). The structure of the analysis of variance is given below

Variation type	Degrees of freedom	Mean Squares	Expected mean squares
Between Cluster	(k-1)	MSB (say)	$\sigma_w^2 + m\sigma_b^2$
Within Cluster	$\sum_{i=1} n_i - k$	MSW (say)	σ_w^2

MSB = Mean square between the clusters; MSW = Mean square within cluster; k = number of clusters; n_i = number of villages in the i^{th} cluster; m is the harmonic mean based on number of villages in each cluster.

Using the mean squares, the estimates of the inter and intra-cluster variances (i.e. $\hat{\sigma}_w^2$ and $\hat{\sigma}_b^2$) are obtained for each cluster. In addition, the ratio of the inter-cluster variance to the total variance is obtained as follows.

$$R^2 = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_w^2}$$

The inter-cluster co-efficient of variation is calculated as follows.

$$CV_b = \frac{\hat{\sigma}_b}{\overline{X}} \times 100$$

Note that, $\overline{X}$ is the general mean for the character.

3. RESULTS AND DISCUSSION

The results are presented below on statistical analyses of data on 5 soil parameters from 20 soil samples taken in each of the 47 villages of Godhara taluka.

Analysis of Variance

Data on the 5 soil parameters from the 20 soil samples taken from each village of Godhara taluka were subjected to ANOVA.

The mean square values and F-test results for each of the five soil parameters are given in Table 1. The data indicated that variation in individual parameters between the villages was highly significant.

The pH values of Godhara taluka varied from 6.3500 to 7.5350, with a mean of 7.2113 (Table 8). The minimum mean value for the village of Mahelol was at par with that of Raysinghpur. The mean values for the remainder of the villages were significantly different from that of Mahelol. Most of these mean values were similar to the maximum mean value of Rampur Jodka.

The minimum, maximum and mean values of EC were 0.2630 ds/m, 0.7630 ds/m and 0.3916 ds/m, respectively (Table 8). A high variation in mean values was observed between the villages with significant differences. The maximum mean value of Rampur Jodka was distinctly higher than that of the other villages.

The OC value of Godhara taluka varied from 0.4550% to 0.6490% with a mean of 0.5894% (Table 8). The maximum mean value of Rinchrota was distinctly higher than that of Harkundi, Raysinghpura, Chhawad, Nani Kantadi, Khajuri Sampa, Pipaliya and Tuwa. The remainder of the villages had mean values at par with that of Rinchrota. The minimum mean value of Tuwa was at par with of the mean values of Pipaliya, Khajuri Sam, Nani Kantad, Chhawad, Raysinghpur and Harkundi.

Regarding P, the minimum, maximum and average values were 10.4390 kg/ha, 29.7000 kg/ha and 20.7747 kg/ha, respectively. The maximum mean value of Rampur Jodka was at par with the mean values of Jitpura, Vavdi Bizar, Vavdi Khurd, Betia, Lilesra, Chanchpur and Chhariya (Table 8).

The K values of Godhara taluka varied from 282.1950 kg/ha to 415.2500 kg/ha, with a mean of 361.7842 kg/ha. The minimum mean value of Mahelol was at par with the mean values of Raysinghpura, Chhabanpur and Nasirpur. The remainder of the villages mean values were distinctly differ that of Mahelol (Table 8).

Multivariate Analysis of Variance

Different multivariate tests, including Pillai's trace, Wilks's lambda, Hotelling's trace and Roy's largest root tests were employed for testing the joint variation of

Table 1. Analysis of variance of different soil parameters

Type of Variation	Degrees of freedom	Mean Squares				
		pH	EC	OC	P	K
Between villages	46	0.828**	0.184**	0.043**	401.91**	12177.86**
Within villages	893	0.431	0.031	0.019	104.69	07208.68

** Significant at 1 % level

Table 2. Characteristics of MANOVA statistics

Effect	Multivariate tests	Value of test statistics	Value of F	Significance
Parameters. EC, pH, OC, P and K	Pillai's Trace	0.637	2.834	0.001
	Wilk's Lambda	0.491	2.960	0.001
	Hotelling's Trace	0.802	3.094	0.001
	Roy's Largest Root	0.411	7.977	0.001

all five soil parameters across the villages. The results are presented in Table 2.

A review of the data shows that the p-values were less than 0.001 indicating highly significant differences between the village-level mean values when all five parameters are considered simultaneously. This implies that when considered together, the five parameters showed heterogeneous values across the villages.

Mahalanobis' D^2

The minimum value of D^2 (0.070) was observed between the villages of Ambali and Odidra; maximum value of D^2 (313.00) was observed between the villages of Rampur Jodka and Tuwa.

Ward's Method

Villages were grouped using Ward's method, which involves squared Euclidean distance. Fig. 1 represents the dendrogram used for cluster analysis.

The value of the fusion coefficient dramatically changed value between the third and fourth clusters. Using the optimum cluster selection procedure described in Aldenderfer and Blashfield (1984), a three-cluster pattern was derived for Godhara taluka. Fig. 2 represents the distribution of villages into

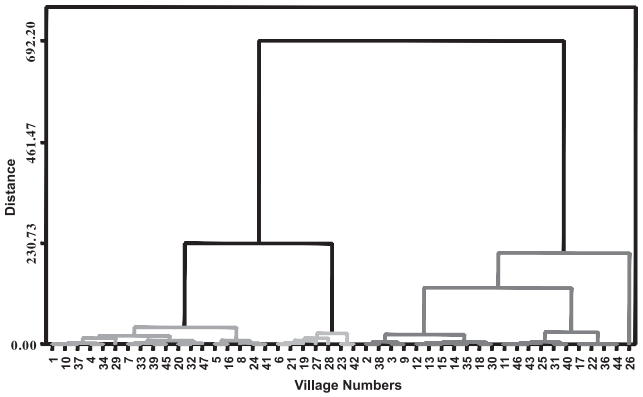

Fig. 1. Ward's Minimum Variance Dendogram formed by the villages of Godhara Taluka

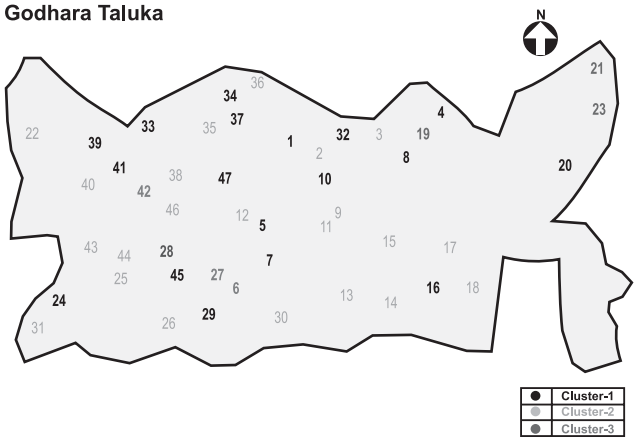

Fig. 2. Distribution of villages in different clusters of Godhara Taluka

different clusters within Godhara taluka. The distribution of villages into three different clusters and the cluster means for the five soil parameters are presented in Table 3.

A maximum number of 22 villages were observed in cluster II followed by 18 in cluster I and 7 in cluster III. The clustering pattern of the villages suggested that geographic diversity may not necessarily be related to village diversity. Rather it may be due to soil heterogeneous patterns.

Among the mean values of pH for different clusters, the value of 7.1906 of cluster II was the lowest value among the clusters, whereas the value of 7.2382 of cluster III was the maximum. The difference between the maximum and minimum mean values was of 0.0476 which indicated that variation was very low.

The EC mean values for the different clusters varied from 0.3221 ds/m (cluster I) to 0.4660 ds/m (cluster II). Very negligible variations were observed; see Table 3. Regarding the mean values of OC the minimum value was observed for cluster III at 0.4973% whereas the maximum was observed for cluster II at 0.6187%. For the mean value of P, cluster

Table 3. Cluster composition and mean values based on Ward's method

Cluster no.	No. of villages	Villages / Mean values					
I	18	Village names	Samali, Sapa, Betia, Ambali, Kankuthambhla, Daruniya, Mahuliya, Orwada, Kalthana, Mahelol, Nadisar, Odidra, Dhanitra, Welvad, Ratanpur, Moryo, Raanipur, Vinjol				
		Mean values	pH	EC	OC	P	K
			7.2382	0.3221	0.5895	22.4010	360.2702
II	22	Village names	Chhabanpur, Nasirpur, Jafrabaad, Bhamaiya, Vavdi Khurd, Lilesra, Dayal, Vavdi Buzarg, Aankadia, Chhariya, Juni Dhari, Chanchpur, Rampur Jodka, Jitpura, Asaardi, Karsana, Rinchrota, Kakkanpur, Moti Kantadi, Ghunsar, Bhima, Veganpur				
		Mean values	pH	EC	OC	P	K
			7.1906	0.4660	0.6187	18.4700	362.7104
III	7	Village names	Nani Kantadi, Khajuri Sampa, Chhawad, Pipaliya, Raysinghpura, Harkundi, Tuwa				
		Mean values	pH	EC	OC	P	K
			7.2072	0.3368	0.4973	23.8357	362.7664

II showed the lowest value at 18.4700 kg/ha whereas the mean value of 23.8357 kg/ha was observed for cluster III. The mean values for K varied from 360.2702 kg/ha (cluster I) to 363.7104 kg/ha (cluster II).

Inter- and Intra-Cluster Distance

Intra- and inter-cluster distances were computed for the three clusters. The data are presented in Table 4.

Table 4. Mean intra- and inter-clusters D^2 values obtained using Ward's method

Cluster	I	II	III
I	**5.77**	36.68	29.20
II		**19.55**	**71.68**
III			**8.34**

* Diagonal values indicate intra-cluster distance

Intra-cluster distance measured in terms of D^2 values ranged from 5.77 in cluster I to 19.55 in cluster II. Meanwhile, inter-cluster distance in terms of D^2 values varied from 36.68 between clusters I and II to 71.68 between clusters II and III.

Inter- and Intra-Cluster Variance

The analysis of variance for each of the five soil parameters was carried out using means of the 47 villages in the three different clusters. To identify which parameter most determined the formation of the three clusters, two benchmarks were used, including R^2 (i.e., the ratio of inter-cluster variances to total variance) and CV_b (i.e., inter-cluster co-efficient of variation in per cent). These values were estimated for each of the five soil parameters. The data are presented in Table 5.

Table 5. Ratio of inter-cluster variances to the total variance (R^2) and inter-cluster coefficient of variation (CV_b)

Sr. No.	Parameters	R^2	CV_b (%)
1	pH	0.00	00.00
2	EC	0.67	24.19
3	OC	0.87	09.51
4	P	0.33	13.67
5	K	0.00	00.00

A maximum R^2 value of 0.87 was obtained for OC, 0.67 for EC, 0.33 for P and 0 for pH and K. The result indicated that the formation of clusters based on these five parameters was largely due to variations in OC, EC and P values. The maximum CV_b value was 24.19% for EC, which was followed by 13.67% for P and 9.51% for OC. The values for pH and K were both 0%. Thus, based on R^2 and CV_b variations in OC, EC and P contributed most to cluster formation.

k-Means Clustering

Villages were also grouped based on all the soil parameters using the k-means clustering method, with k=3. A maximum number of 24 villages were observed in cluster II, which included all 21 villages that appeared in cluster II under Ward's method except for the village of Rampur Jodka. In addition, the villages of Nani Kantadi, Khajuri Sampa, Chhawad, Pipaliya, Raysinghpura, Harkundi and Tuwa now appeared together in cluster I (Table 6).

An analysis of variance based on the k-means clustering method revealed that the cluster formation was largely determined by variations in OC, EC and P values (Table 7).

4. CONCLUSION

The variations in all five individual soil parameters among the 47 villages under study were highly significant, indicating that the selected variables significantly differed among the villages. The variability within villages was non-significant due to homogeneity within the villages. A MANOVA test revealed significant variability between the villages of the Godhara taluka when all the five soil parameters were considered simultaneously. Cluster formation was largely determined by the variations in OC, EC and P values, as confirmed by Ward's method as well as k-means clustering. Three clusters were formed such that there was homogeneity within the clusters and

Table 6. Cluster composition based on k-means clustering

Cluster no.	No. of villages	Villages / Mean values					
I	22	Village names	Samali, Sapa, Betia, Ambali, Daruniya, Mahuliya, Orwada, Mahelol, Nadisar, Odidra, Dhanitra, Welvad, Ratanpur, Raanipur, Vinjol, Nani Kantadi, Khajuri Sampa, Chhawad, Pipaliya, Raysinghpura, Harkundi, Tuwa				
		Mean values	pH	EC	OC	P	K
			7.2217	0.3176	0.5595	22.7758	362.5418
II	24	Village names	Chhabanpur, Nasirpur, Jafrabaad, Bhamaiya, Vavdi Khurd, Lilesra, Dayal, Vavdi Buzarg, Aankadia, Chhariya, Juni Dhari, Chanchpur, Jitpura, Asaardi, Karsana, Rinchrota, Kakkanpur, Moti Kantadi, Ghunsar, Bhima, Veganpur, Kankuthambhla, Moryo, Kalthana				
		Mean values	pH	EC	OC	P	K
			7.1883	0.4440	0.6151	19.3709	360.9141
III	1	Village name	Rampur Jodka				
		Mean values	pH	EC	OC	P	K
			7.5350	0.7630	0.6330	10.44	366.0000

Table 7. ANOVA table based on k-means clustering

Parameters	Cluster		Error		F	Significance
	Mean Square	d.f.	Mean Square	d.f.		
PH	0.320	2	0.218	44	1.478	0.239
EC	174.426	2	2.448	44	71.240	0.000
OC	54.424	2	3.752	44	14.507	0.000
P	0.006	2	0.001	44	4.343	0.019
K	0.000	2	0.000	44	0.896	0.415

Table 8. Mean values of soil parameters for villages in Godhara Taluka

COUNT	VID	Village Name	pH	EC	OC	P	K
1	V10036	Samali	7.1250	0.3145	0.5670	22.60	328.76
2	V10037	Chhabanpur	7.0650	0.4575	0.5810	18.40	322.06
3	V10039	Nasirpur	7.2250	0.4940	0.5680	19.10	322.70
4	V10040	Sapa	7.4850	0.2855	0.5725	25.05	361.76
5	V10047	Betia	7.0950	0.3410	0.5815	15.75	370.50
6	V10048	Nani Kantadi	7.3700	0.2980	0.5020	24.20	393.75
7	V10049	Ambali	7.3850	0.3075	0.5950	24.75	403.00
8	V10054	Kankuthambhla	7.2500	0.3775	0.6085	19.65	342.31
9	V10056	Jafrabaad	7.3150	0.4870	0.5780	20.80	377.50
10	V10057	Daruniya	7.2000	0.3160	0.5725	17.40	382.75
11	V10058	Bhamaiya	7.1400	0.3770	0.6320	20.85	377.75
12	V10059	Vavdi Khurd	7.1050	0.5215	0.6145	13.85	361.50
13	V10062	Lilesra	6.9950	0.5090	0.6065	15.75	341.50
14	V10064	Dayal	7.2150	0.4945	0.6020	17.70	377.50
15	V10065	Vavdi Buzarg	7.1150	0.5095	0.6050	13.20	381.50
16	V10068	Mahuliya	7.1900	0.3595	0.5670	17.20	350.01
17	V10069	Aankadia	7.1950	0.4500	0.6350	19.70	346.56
18	V10070	Chhariya	7.1200	0.5155	0.6310	16.15	361.06
19	V10075	Khajuri Sampa	7.1700	0.3570	0.4945	18.45	356.02
20	V10079	Orwada	7.3850	0.3450	0.6060	26.20	362.50
21	V10086	Chhawad	7.3000	0.3200	0.5045	29.70	341.50
22	V10087	Juni Dhari	7.2350	0.4420	0.6340	19.40	376.50
23	V10089	Pipaliya	7.3150	0.3335	0.4640	29.20	356.50
24	V10090	Kalthana	7.3650	0.4015	0.5945	22.50	349.25
25	V10091	Chanchpur	7.1500	0.4085	0.6340	15.78	365.75
26	V10092	Rampur Jodka	7.5350	0.7630	0.6330	10.44	366.00
27	V10096	Raysinghpura	6.5655	0.3400	0.5285	20.00	311.90
28	V10097	Harkundi	7.3350	0.3995	0.5325	19.65	415.25
29	V10098	Mahelol	6.3500	0.2630	0.5755	19.85	282.20
30	V10099	Jitpura	7.0950	0.5420	0.6235	10.56	377.50
31	V10107	Asaardi	7.2850	0.3385	0.6325	24.00	340.75
32	V10108	Nadisar	7.3550	0.3360	0.6085	27.85	359.55
33	V10111	Odidra	7.3900	0.3010	0.5980	23.80	365.50

COUNT	VID	Village Name	pH	EC	OC	P	K
34	V10113	Dhanitra	7.4100	0.2810	0.5760	23.85	412.25
35	V10114	Karsana	7.1490	0.4830	0.6045	17.66	366.50
36	V10116	Rinchrota	7.1000	0.4540	0.6490	24.90	354.25
37	V10118	Welvad	7.3200	0.3040	0.5820	22.75	386.00
38	V10119	Kankanpur	7.2200	0.4450	0.5955	22.85	383.25
39	V10120	Ratanpur	7.3850	0.2795	0.6060	24.60	370.00
40	V10123	Moti Kantadi	7.2850	0.3560	0.6410	22.10	375.25
41	V10124	Moryo	7.2450	0.3890	0.5800	26.85	356.75
42	V10125	Tuwa	7.3950	0.3095	0.4550	25.65	364.45
43	V10126	Ghunsar	7.3000	0.3845	0.6285	21.30	368.25
44	V10128	Bhima	7.1400	0.4445	0.6450	20.60	379.00
45	V10130	Raanipur	7.0100	0.2795	0.6010	17.20	336.01
46	V10131	Veganpur	7.2100	0.3750	0.6380	21.25	357.00
47	V10133	Vinjol	7.3421	0.3163	0.6200	25.37	365.79
		Minimum =	**6.3500**	**0.2630**	**0.4550**	**10.44**	**282.20**
		Maximum =	**7.5350**	**0.7630**	**0.6490**	**29.70**	**415.25**
		Average =	**7.2113**	**0.3916**	**0.5894**	**20.7747**	**361.7842**
		CD =	**0.4070**	**0.1090**	**0.0860**	**6.3420**	**41.8370**

heterogeneity between the clusters. These groupings can be used to prepare fertility maps and to develop effective soil improvement programmes. Though distinct clusters were identified in Godhara taluka, geographical closeness among the villages in a given cluster was not evident, indicating that the factors influencing the soil parameters were geographically well dispersed. Ward's method was found suitable to group the villages of the Godhara taluka.

ACKNOWLEDGEMENTS

The authors are thankful to the referee for providing valuable suggestions to improve the quality of the paper.

REFERENCES

Aldenderfer, Mark S. and Blashfield, Roger K. (1984). *Cluster Analysis*. Sage Publications, Newbury Park, California.

Anderberg, M.R. (1973). *Cluster Analysis for Applications*. Academic Press, INC (LONDON) LTD.

Dixit, S.K. (1984). A biometrical study of genetic divergence in macroni wheat under irrigated conditions. Unpublished Ph.D. thesis. Gujarat Agricultural University, Anand Campus, Anand.

Johnson, Richard A. and Wichern, W. Dean (2002). *Applied Multivariate Statistical Analysis*. Prentice-Hall of India Pvt. Ltd, New Delhi.

Rao, C.R. (1952). *Advanced Statistical Methods in Biometrical Research*. John Wiley and Sons, New York, USA, 236-272.