Indian Journal of Agricultural Sciences **81** (8): 756–9, August 2011

Average linkage method for clustering rice (*Oryza sativa*) producing states of India*

AMRIT KUMAR PAUL[1], WASI ALAM[2] and PAL SINGH[3]

Indian Agricultural Statistics Research Institute, New Delhi 110 012

Received: 14 May 2009; Revised accepted: 23 May 2011

Key words: Average linkage, Rice production, Single linkage, Squared euclidean distance

Rice (*oryza sativa*) is India's preeminent crop, and is the staple food of the people of the eastern and southern parts of the country. To know the trend of rice production of India, we have considered rice production of 15 years for 34 states (including union territory).

Efforts were made to group the similar rice producing states by using cluster analysis (Hair *et al.* 1998; Johnson and Wichern 1996; Timm 2002) statistical technique, this is the unique statistical technique where there are many alternative options of clustering methods as well as similarity or distance measures for solving the same kind of problem. Average linkage and single linkage methods of hierarchical cluster analysis (Tseng and Wong 2005, Romesburg 1984) have been used for clustering the rice producing states of India into valid clusters. Squared Euclidean distance has been considered as distance measure. It has been shown that in spite of using the same distance measure for both the clustering methods, performance of single linkage method is poor than average linkage method. An effort has been made to demonstrate the overall impact of the choice of clustering methods in getting the valid clusters. The obtained valid clusters of rice producing states have been further studied. Cluster-wise trend of rice production level has been depicted in figures (1–4).

State-wise total rice production ('000 tonnes) of 34 states of India for 15 years (from 1990 to 2004) was taken from *Agricultural Statistics at a Glance* (*1990 to 2004*), Directorate of Economics and Statistics, Department of Agricultural and Co-operation, Ministry of Agriculture, Government of India.

Average linkage computes the distance between subgroups at each step as the average of the distances between the two subgroups. Let $s \in S$ and $m \in M$, where S and M are two clusters, distances between S and M are calculated by using the rule

$$d_{(S)(M)} = \frac{\sum_S \sum_M d_{sm}}{n_S S_M}$$

where, n_s and S_M are the number of objects in each cluster. Distance in step3 are replaced by an average of $n_S S_M$ distances between all pairs of elements and $m \in M$.

Squared euclidean distance is the square of the standard euclidean distance and it is used to place progressively greater weight on objects that are further apart. This distance is computed as distance $(x,y) = {}_i (x_i - y_i)^2$, where, i = 1,2,…8 for fifteen years.

Descriptive statistics given in Table 1 can give only rough idea about the rice production level of different states of India. These descriptive statistics is based on 15 years data, mean value is drastically affected in presence of a single extreme value. Hence to reduce the impact of extreme values or outliers in calculation of NAAV, 5% trimmed mean has been used for calculating the NAAV. For grouping the rice producing states into valid clusters average linkage method of agglomerative hierarchical clustering algorithm has been applied. According to the rescaled dendrogram obtained (Fig 1) through the average linkage method over squred Euclidean distance, all the thirty four states have been classified into four distinct and valid clusters. In the first cluster, there are 22 states whose production levels fall below the national average (NAAV), 2377.51 ('000 tonnes), viz Chandigarh, Daman and Diu, Delhi, Sikkim, Dadra and Nagar Haveli, Andaman, Puducherry, Mizoram, Himachal Pradesh, Arunachal Pradesh, Goa,Meghalaya, Rajasthan, Nagaland, Manipur, Tripura, Jammu and Kashmir, Uttaranchal, Gujarat, Kerala, Madhya Pradesh and Jharkhand, second cluster consist of six states, viz Bihar, Maharashtra, Haryana, Asom, Karnataka and Chhattisgarh, whose average production levels are closely above than the national average (NAAV). Third cluster consist of three states namely Odisha, Tamil Nadu and Punjab whose average production levels are above than that of states under cluster

*Short note

[1]Senior Scientist (e mail: pal@iasri.res.in); [2]Scientist (SS) (e mail: wasi@iasri.res.in), [3]Scientist (SS) (e mail: ps@iasri.res.in)

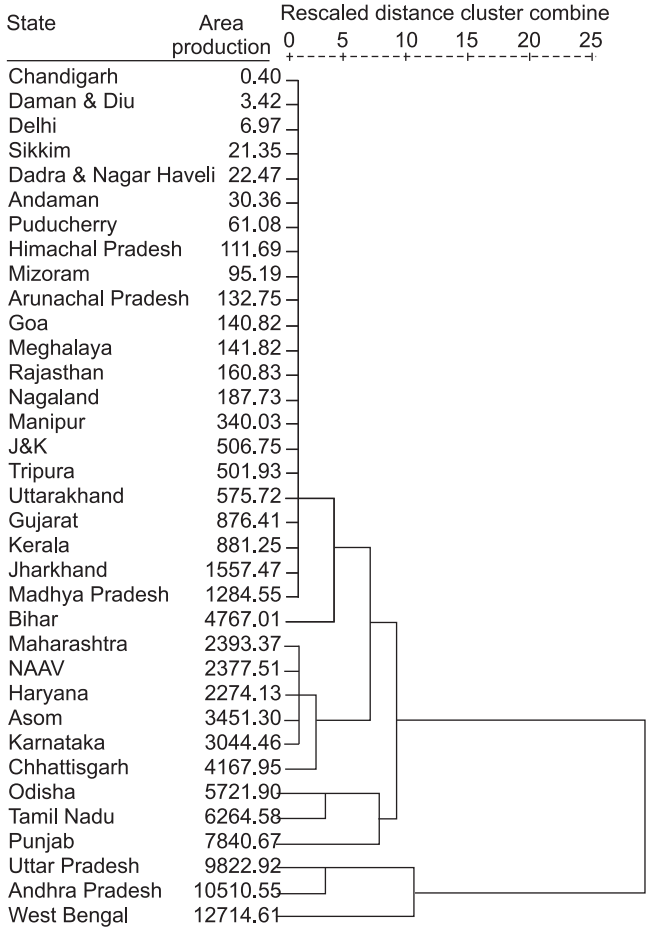

Fig 1 Dendrogram using average linkage (within group) rescaled distance cluster combine

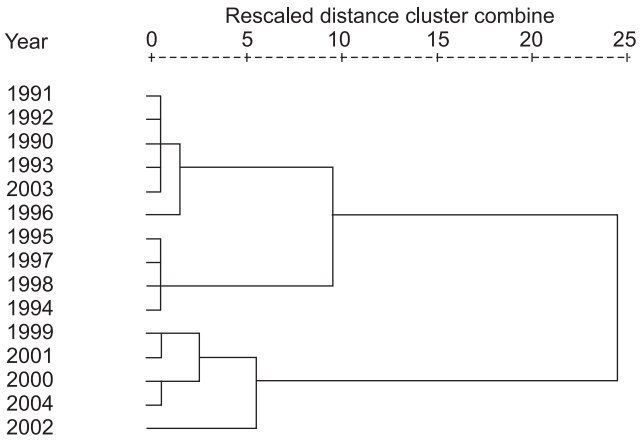

Fig 2 Hierarchical cluster analysis clustering of years based on total production of India. Dendrogram using average linkage (within group)

number two and the fourth cluster consist of three states, viz Andhra Pradesh, Uttar Pradesh and West Bengal average rice production levels are significantly above the NAAV.

On the basis of total rice production of India all the 15 years have been grouped in terms of total rice production and is depicted in Fig 2. First cluster consist of 1990, 1991, 1992, 1993 and 2003, second cluster consist of 1994 to 1996, third cluster consist of 1999,2000,2001,2004 and fourth cluster consists of only 2002. On the basis of rice production of 34 states all the 15 years have been grouped in hierarchical fashion and has been depicted in Fig 3. Figs 2 and 3 show that rice production in 2002 was very different as compared to other years which is further revealed in the Fig 4. Major rice producing state like West Bengal which has significantly above level than NAAV, has got the highest production in 2002 (Fig 4), similarly others and even it is true for Asom which has very close production level to NAAV.

Clusters obtained through average linkage method has been considered for demonstrating the trend of some of the rice-producing states over along with NAAV and have been depicted in Fig 4. Which shows that West Bengal has highest production level and there is almost continuous rise trend across all the years. West Bengal is the only state, which has far above production level than the national average and has continuous rise in the rice production over the 15 years.

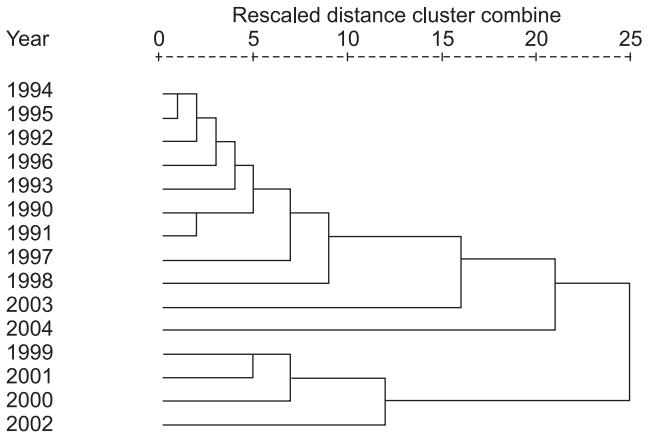

Fig 3 Clustering of years based on 15 years total rice production of all the 34 statges

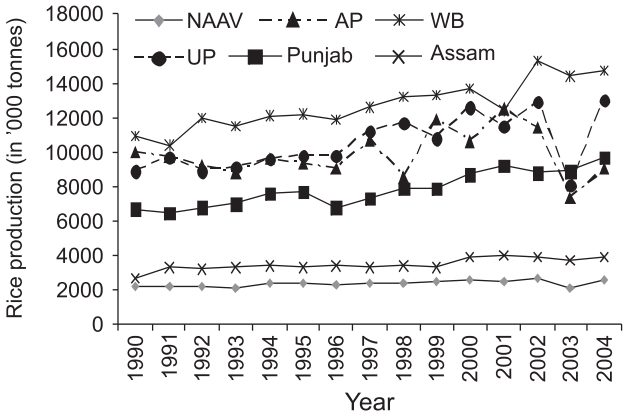

Fig 4 Total rice production ('000 tonnes) vs year

Table 1 State-wise average rice production ('000 tonnes)

State	N	Minimum	Maximum	Mean	Std. deviation
Chandigarh	15	0.25	0.70	0.40	0.13
Daman and Diu	15	2.90	4.50	3.42	0.57
Delhi	15	1.80	19.40	6.97	4.73
Sikkim	15	12.90	25.30	21.35	2.64
Dadra	15	17.00	33.20	22.47	4.31
Andaman	15	26.20	33.20	30.36	2.20
Puducherry	15	51.46	68.20	61.08	4.81
Mizoram	15	59.20	113.30	95.19	17.94
Himachal Pradesh	15	85.70	137.40	111.69	12.78
Arunachal Pradesh	15	105.80	154.60	132.75	13.89
Goa	15	126.50	170.70	140.83	10.87
Meghalaya	15	111.50	200.70	141.82	31.01
Rajasthan	15	67.90	252.60	160.83	42.42
Nagaland	15	150.00	237.30	187.73	28.63
Manipur	15	245.12	387.30	340.03	42.58
Tripura	15	413.90	602.30	501.93	51.23
Jammu and Kashmir	15	391.50	589.10	506.75	64.66
Uttaranchal	15	483.00	621.50	575.72	35.13
Gujarat	15	472.70	1 277.00	876.41	196.59
Kerala	15	624.00	1 142.14	881.25	171.18
Madhya Pradesh	15	937.90	1 750.00	1 284.55	266.37
Jharkhand	15	973.00	2 310.00	1 557.47	355.07
Haryana	15	1 750.00	2 793.00	2 274.13	377.75
NAAV	15	2 112.36	2 745.29	2 377.51	197.65
Maharash	15	1 850.00	2 839.00	2 393.37	265.32
Karnataka	15	2 269.26	3 846.74	3 044.46	493.01
Asom	15	2 671.65	3 998.50	3 451.30	351.11
Chhattisgarh	15	2 550.00	5 410.40	4 167.95	869.04
Bihar	15	2 606.00	6 126.12	4 767.01	850.67
Odisha	15	3 240.00	7 148.40	5 721.90	1 063.22
Tamil Nadu	15	3 222.78	8 141.40	6 264.58	1 390.33
Punjab	15	6 535.00	9 656.00	7 840.67	999.44
Andhra Pradesh	15	7 326.79	12 458.00	9 822.92	1 368.14
Uttar Pradesh	15	8 116.59	13 018.80	10 510.55	1 574.39
West Bengal	15	10 436.50	15 256.70	12 714.61	1 372.55
India total Mean	15	71 820.10	93 340.00	80 835.49	6 719.94

Source: *Agricultural Statistics at a Glance* (*1990 to 2004*)

Production trend and level of Andhra Pradesh and Uttar Pradesh are very much similar and close to each other. In 2003, there is a sharp decline in the production level of Andhra Pradesh, Uttar Pradesh and West Bengal. In Punjab, there is no decline in production over the 15 years even in 2003, when major producing states demonstrated a major decline in production. Production level of Asom and NAAV are some what stable over the fifteen years. States under cluster-1 are lowest producing states.

SUMMARY

This study can provide an opportunity to discover more precisely the clusters of least rice-producing states as well as clusters of moderate rice-producing states, whose production level marginally fall below the national average. This can assist in exploring the reasons of small production in a particular cluster of geographical locations. Cluster of the geographical locations of high productivity can further be studied for improving the low rice producing states. Application of average linkage clustering method is more effective statistical technique in identification of homogeneous group of geographical locations producing the similar rice production level.

REFERENCES

Agricultural Statistics at a Glance (*1990 to 2004*). Directorate of Economics and Statistics, Department of Agricultural and Co-

operation, Ministry of Agriculture, Government of India.

Hair J J F, Anderson R E, Tatham R L and Black W C. 1998. *Multivariate Data Analysis*, 420 pp. Pearson Education, Singapore, Pvt. Ltd.

Johnson R A and Wichern D W. 1996. *Applied Multivariate Statistical Analysis*, 573 pp. Prentice Hall of India Pvt Ltd, New Delhi.

Romesburg H C. 1984. *Cluster Analysis for Researchers*. Life time learning publications.

Timm N H. 2002. *Applied Multivariate Analysis*, 515 pp. Springer-Verlag, New York.

Tseng G C and Wong W H. 2005. Tight clustering: A resampling based approach for identifying stable and tight patterns in data. *Biometrics* **61**: 10–16.