# An Empirical Investigation on Classical Clustering Methods

S D Wahi[*], Sukanta Dash[**] and A R Rao[***]

Five classical clustering methods: four hierarchical—single linkage, average-between linkage, average-within linkage, Wards—and one non-hierarchical—k-means—using five different distance measures: squared Euclidean, city block, Chebychev's, Pearson correlation and Minkowski have been compared on the basis of simulated multivariate data on paddy crop genotypes. The performance of different clustering methods was compared based on the average percentage probability of misclassification and its standard error. The performance of different hierarchical clustering methods varied with distance measures used and it was found that squared Euclidean performed best among the five distances followed by city block distance in majority of cases. Among the five methods, the Ward's method performed best with least average percentage probability of misclassification followed by non-hierarchical k-means method irrespective of the sample size. Among the different distance measures used under hierarchical clustering methods, the squared Euclidean distance showed least average percentage probability of misclassification followed by city block distance.

*Keywords:* Cluster analysis, Rice, Hierarchical methods, Non-hierarchical method, Distance measures

## Introduction

The summarization of large quantities of multivariate data is being increasingly practiced in various branches of agricultural science. A number of multivariate statistical techniques, namely, cluster analysis, principal components analysis, factor analysis are being widely used for classification purposes. One of the basic problems faced by the plant breeders is to classify large number of genotypes/lines into fewer manageable homogeneous groups/clusters. There are large number of clustering methods and dissimilarity measures available in literature for making homogeneous groups. One of the main problems faced by the breeder is to choose a suitable method of clustering and dissimilarity measure among the different methods and dissimilarity measures available in literature. There is hardly any information available in literature on the performance of these clustering methods and dissimilarity measures. Researchers commonly use UPGMA (Unweighted Pair-Group Method using Arithmetic Averages) and Ward's method followed by SLINK (Single Linkage) and CLINK (Complete Linkage) among the existing clustering methods. According to Blashfield (1976) UPGMA, Ward's and SLINK account for 3/4th of

[*] Principal Scientist, Indian Agricultural Statistics Research Institute, New Delhi, India; and the corresponding author. E-mail: sdwahi.iasri.res.in

[**] Research Scholar, Indian Agricultural Statistics Research Institute, New Delhi, India. E-mail: sukanta.iasri@gmail.com

[***] Senior Scientist, Indian Agricultural Statistics Research Institute, New Delhi, India. E-mail: arrao@iasri.res.in

the published work which used cluster analysis technique. The lesser used cluste methods, which appear occasionally in applications are WPGMA (Weighted Pair-Gr Method using Arithmetic Averages) method, the centroid method and the flexible met (Sneath and Sokal, 1973). Lin (1982), Ramey and Rosielle (1983), Wahi and Kher (19 promoted the application of clustering techniques to group the genotypes environments but the number of clusters obtained from these methods are not uni because of unrepresentativeness of the clustering groups obtained through diffe clustering procedures. k-means method requires prior knowledge of the numbe clusters but unfortunately in the case of unsupervised classification usually there is prior idea about the number of clusters. On the contrary, hierarchical clustering meth do not require a prior knowledge of number of clusters, which is a definite advant over k-means method.

In the present investigation five clustering methods: four hierarchical—single linka average-between linkage, average-within linkage, Wards—and one non-hierarchica k-means clustering are empirically compared using simulated multivariate data of pa crop. The performance of the hierarchical clustering procedures based on five dista measures, namely, squared Euclidean, city block, Chebychev's, Pearson correlation Minkowski is also assessed.

## Materials and Methods

The commonly used classical methods of clustering fall into two general categor hierarchical and non-hierarchical (Johnson and Wichern, 2006). The following five cluste methods are considered in the present study:

1. Single linkage (minimum distance or nearest neighbor);
2. Complete linkage (maximum distance or farthest neighbor);
3. Average linkage (average distances);
4. Ward's Hierarchical Clustering Methods; and
5. Non-Hierarchical Clustering Method—k-means Clustering.

The five distance measures that have been utilized for hierarchical clustering in present investigation are as follows:

**Squared Euclidean Distance:** The squared Euclidean distance between two $p$-dimensio observations $x=[x_1,x_2,\dots,x_p]'$ and $y=[y_1,y_2,\dots,y_p]'$ is $d(x,y)=\sum_i (x_i-y_i)^2$

**City block or Manhattan Distance:** This distance is simply the average difference acr dimensions. The city block distance is computed as:

$$d(x,y)=\sum_i |x_i - y_i|$$

**Chebychev Distance:** The Chebychev distance is computed as:

$$d(x,y)=Max.|x_i - y_i|$$

**Minkowski Distance:** This distance is defined as:

$$d(x,y)=\left[\sum_i |x_i - y_i|^m\right]^{1/m}$$

**Pearson Correlation:** This metric is defined as:

$$r(x,y)=\frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

where $\sigma_{xy}$ is the covariance between two *p*-dimensional observation $\sigma_x$ and $\sigma_y$ and are the standard deviations of *x* and *y* respectively.

The performance of different clustering methods and distance measures have been compared by using the average percentage probability of misclassification obtained from the simulated samples. Further, the consistency of different methods has been adjudged on the basis of standard error of the average percentage probability of misclassification.

## Data used for Simulation

Seventy five rice genotypes consisting of 25 tall, 25 medium and 25 dwarf on nine traits like tiller number per plant, plant height in cm, panicle length in cm, panicle weight in g, thousand grain weight in g, biomass per plant in g, harvest index in percentage, straw yield per plant in g and grain yield per pant in g are taken from Genetics Division of IARI, New Delhi and are used for simulation of multivariate data of different sample sizes.

## Results and Discussion

Four homogeneous clusters based on multivariate data on 75 rice genotypes were obtained from the results of different clustering procedures on consensus basis. The mean vectors and dispersion matrices of the four clusters so obtained were used as population parameters for simulation purposes. Four different hierarchical methods of clustering given in material and methods, using five different distance measures and one non-hierarchical method of clustering, i.e., k-means are compared using four simulated multivariate normal populations with different mean vectors

and dispersion matrices. The performance of these methods was comp
basis of average probability of misclassification and its standard error of
20 different simulated samples. Further, the consistency of these methods
on the basis of standard errors of these probabilities. The samples of
data for four clusters of small (= 30), moderate (= 60) and large (= 150)
are simulated. The average probabilities of misclassification along with
errors for different methods of clustering are given in Table 1. The result
among the five clustering methods the average percentage of miscl
is least for Ward's method followed by k-means method irrespective of
size. The average percentage of probability of misclassification of a l
clustering method varied widely with the distance measure used. Ha
for Ward's method as the best method, the probability of misclass

**Table 1: Average Percentage Probability of Misclassification (*P*) and its Standar for Different Methods of Clustering and Distance Measures**

| | | | Clustering Method | | | |
|---|---|---|---|---|---|---|
| Sample Size | Distance | | Slink | Complete | Average | Ward's |
| Small | Euclidean | *P* | 30.38 | 14.13 | 8.63 | 4.88 |
| | | SE | 2.53 | 1.32 | 0.64 | 0.42 |
| | City block | *P* | 31.50 | 14.75 | 9.13 | 5.00 |
| | | SE | 1.73 | 1.46 | 0.55 | 0.41 |
| | Chebychev | *P* | 32.38 | 14.88 | 9.75 | 5.88 |
| | | SE | 1.53 | 1.00 | 0.75 | 0.49 |
| | Pearson | *P* | 32.88 | 14.25 | 9.13 | 6.00 |
| | Correlation | SE | 1.57 | 1.01 | 0.55 | 0.46 |
| | Minkowski | *P* | 31.50 | 15.13 | 9.25 | 6.13 |
| | | SE | 2.20 | 0.86 | 0.60 | 0.61 |
| Medium | Euclidean | *P* | 28.46 | 13.88 | 7.63 | 4.75 |
| | | SE | 2.54 | 1.05 | 0.65 | 0.49 |
| | City block | *P* | 29.09 | 14.56 | 7.75 | 5.20 |
| | | SE | 2.43 | 1.17 | 0.63 | 0.52 |
| | Chebychev | *P* | 29.06 | 14.31 | 7.94 | 5.63 |
| | | SE | 2.55 | 1.22 | 0.64 | 0.50 |
| | Pearson | *P* | 29.09 | 14.63 | 7.69 | 5.50 |
| | Correlation | SE | 2.51 | 1.08 | 0.70 | 0.37 |

Table 1 (Cont.)

| Sample Size | Clustering Method | | | | | | |
|---|---|---|---|---|---|---|---|
| | Distance | | Slink | Complete | Average | Ward's | k-means* |
| | Minkowski | *P* | 31.83 | 14.31 | 7.63 | 5.56 | |
| | | SE | 1.94 | 1.31 | 0.69 | 0.53 | |
| Large | Euclidean | *P* | 24.45 | 11.43 | 7.05 | 4.08 | 4.83 |
| | | SE | 2.07 | 0.85 | 0.46 | 0.34 | 0.34 |
| | City block | *P* | 24.53 | 11.83 | 8.58 | 4.20 | |
| | | SE | 2.19 | 0.76 | 0.81 | 0.32 | |
| | Chebychev | *P* | 26.80 | 12.75 | 8.65 | 5.30 | |
| | | SE | 1.94 | 1.15 | 0.62 | 0.34 | |
| | Pearson Correlation | *P* | 28.45 | 14.80 | 8.48 | 5.58 | |
| | | SE | 1.66 | 1.13 | 0.51 | 0.46 | |
| | Minkowski | *P* | 27.48 | 14.90 | 7.78 | 4.86 | |
| | | SE | 1.74 | 1.17 | 0.46 | 0.33 | |

**Note:** * k-means method is based on nearest centroid (mean).

least (4.88) with Euclidean distance followed by city block distance (5.00) for small sample size. Similarly, probability of misclassification is least (4.75) with Euclidean distance followed by city block distance (5.20) for medium sample size and least (4.08) with Euclidean distance followed by city block (4.20) for large sample size. The results further show that between Ward's and k-means methods the standard error of average percentage of probability of misclassification is least in k-means method under small sample size (= 0.40) and under medium sample size (= 0.45) whereas for large sample size the standard errors of both the methods are same. ❃

## References

1. Blashfield R K (1976), "Questionnaire on Cluster Analysis Software", *Classification Society Bulletin*, Vol. 3, No. 4, pp. 25-42.

2. Johnson R A and Wichern D W (2006), *Applied Multivariate Statistical Analysis*, 5th Edition, London Inc., Pearson Prentice Hall.

3. Lin C S (1982), "Grouping Genotypes by a Cluster Method D Genotype-Environment Interaction Mean Square", *Theoretical an* Vol. 62, pp. 277-280.

4. Ramey T B and Rosielle A A (1983), "HASS Cluster Analysis: A Grouping Genotypes or Environments in Plant Breeding", *Theo. Genetics*, Vol. 66, pp. 131-133.

5. Sneath P H A and Sokal R R (1973), *Numerical Taxonomy*, Freeman

6. Wahi S D and Kher K K (1991), "A Comparison of Clustering Pro Multiple Traits in Gerbera and Dahlia", *Indian J. Genetics.*, pp. 335-341.

*Reference #* ι