Evaluation of different clustering techniques in classifying the vegetable growing panchayats of Ernakulam district, Kerala

ABSTRACT

The goal of this study was to evaluate different clustering techniques in classifying the vegetable growing locations of Ernakulam (EKM) district of Kerala so that same nutrient recommendation can be prescribed for panchayats coming under the same cluster. Hierarchical clustering (HC) and *K* –means clustering were performed to group the panchayats based on soil fertility status and thereafter comparison of various clustering procedures was done using Davies – Bouldin (DB) index. Different dissimilarity measures- Euclidean, squared Euclidean, Chebychev distance and Mahalanobis D^2 were determined and single linkage, complete linkage and average linkage methods were adopted under these measures. The results revealed that Mahalanobis D^2 was the better clustering procedure with seven clusters (DB index: 0.120) followed by average linkage method under Euclidean distance (DB index: 0.306) with seven clusters. Manjapra and Keerampara panchayats remained as individual clusters. Keerampara had strongly acidic soils (pH -5.17) with high available Mg (158 mg kg-1) while Manjapra soils had low Mg availability (19 mg kg-1) and high S content (57 mg kg-1). Kakkad, Kalady and Vengoor came under cluster I which possessed approximately same EC (0.15-0.19 dS m-1), OC (2-2.4%) and Mg (71-73 mg kg-1) content. Chengamanadu and Vengola came under cluster III while Ayyampuzha and Mudakkuzha belong to cluster IV.

Keywords: *Hierarchical clustering, Davies – Bouldin index, Mahalanobis D2, Euclidean distance, Average linkage*.+

1. INTRODUCTION

Kerala, the God's own country is blessed with biological diversity and soil heterogenity. Ernakulam is a district situated in the central region of Kerala and rice is the major crop cultivated in the wet lands of the district. In addition to rice, vegetables, pineapple and nutmeg are also cultivated in EKM. Plant growth is highly dependent on the fertility status of soil prevailing in the region [1]. Plants need primary nutrients (Nitrogen (N), Phosphorus (P), Potassium (K)), secondary nutrients (Calcium (Ca), Magnesium (Mg), Sulphur (S)) and micro nutrients (Boron (B), Iron (Fe), Manganese (Mn), Zinc (Zn), Copper (Cu)) in adequate quantities for growth and development and they are absorbed from the soil. Other soil parameters viz. pH and Electrical conductivity (EC) also affect plant growth and nutrient availability [2].

Systematic testing of soil nutrients including the micro nutrients and prescribing recommendation is essential for soil health management. The soil test data available for each panchayat can be used for classifying them into various groups based on similarity in soil parameters. Cluster analysis, one of the multivariate methods is suitable in classifying objects based on similarity measures [3]. Clustering of panchayats based on their soil fertility status can be accomplished by means of cluster analysis [4]. Panchayats coming under the same cluster have similar soil characteristics and that under different clusters have dissimilar soil properties. Different clustering procedures are available which vary according to the distance measures selected for clustering. Same recommendations can be given to those panchayats which come under the same cluster.

Cluster analysis is commonly applied in the field of plant breeding and genetics where varieties/genotypes can be grouped based on their quantitative characters [5]. According to [6], results of cluster analysis changes when different methods of clustering are used. Cluster analysis

can be used to partition sites based on soil characteristics [7]. The land use effect on soil chemical and microbial properties can be determined with the use of cluster analysis [8]. Application of cluster analysis on soil geochemical data helped to determine the spatial distribution of elements [9]. Evaluation of different clustering techniques is possible with a measure; Davies- Bouldin index which identifies the cluster that best fit the data [10].

This study is an attempt to classify the panchayats of Ernakulam district based on the soil fertility status. Among the various clustering procedures, clusters that best fit the data are also identified.

2. MATERIALS AND METHODS

The present study is based on the data on thirteen soil fertility parameters of different vegetable growing locations (panchayats) of Ernakulam district of Kerala. Soil samples collected from different panchayats of Ernakulam district during the year 2016-2017 analysed by Department of Soil science & Agricultural Chemistry, College of Agriculture, Vellayani, Kerala and the data maintained was utilized for the present study. Sample collection was done by the farmers themselves from their own vegetable growing plots at a depth of 15 cm using spade. From each sampling plot, about 10-15 random samples were collected, mixed and reduced to 0.5 Kg by the method of quartering. It was observed that panchayats show variation in soil properties with respect to cropping patterns and cultivation practices [11].

The data on thirteen soil fertility parameters of 17 panchayats of Ernakulam viz., electro chemical parameters (pH and Electrical conductivity (EC)), Oxidisable Organic Carbon (OC), Phosphorus (P), Potassium (K)), secondary nutrients (Calcium (Ca), Magnesium (Mg), Sulphur (S)) and micro nutrients (Boron (B), Iron (Fe), Manganese (Mn), Zinc (Zn), Copper (Cu)) were available. Each panchayat is having different sample sizes and altogether sample size comes around 583.

2.1 Cluster analysis (CA)

Cluster analysis is a multivariate technique used to group individuals or objects based on their several characteristics [12]. There should be homogeneity within groups or clusters and heterogeneity between groups. First we measure the distance between objects based on their multiple characters which is otherwise called as similarity or dissimilarity measures. Based on the similarity/dissimilarity measures, clusters are formed later.

2.1.1 Distance measures

Euclidean distance, squared Euclidean distance, Chebychev distance and Mahalanobis D^2 are the distance measures used for the present work.

2.1.1.1 Euclidean distance

It is the geometrical distance between two objects in the multidimensional space. It is calculated as,

$$d(x, y) = \sqrt{(X - Y)'(X - Y)}$$

where X and Y are the 'p' dimensional vector of observations and $X = (X_1, X_2, \ldots X_p)$ and $Y = (Y_1, Y_2, \ldots Y_p)$. Euclidean distance is one of the most commonly used distance measures.

2.1.1.2 Squared Euclidean distance

It is the square of Euclidean distance and is used to put weights for those objects which are farther apart.

$$E_{ij} = \sum_{k=1}^{p}(X_{ik} - X_{jk})^2$$

2.1.1.3 Mahalanobis D^2 statistics

A measure for group distance based on multiple characters was given by Mahalanobis, 1936. With x_1, x_2, x_3…., x_p as multiple measurements available on each individual and d_1, d_2, d_3,…, d_p as $\bar{x}_1^1 - \bar{x}_1^2$, $\bar{x}_2^1 - \bar{x}_2^2$,……., $\bar{x}_p^1 - \bar{x}_p^2$, respectively, being the difference in the means of two populations, Mahalanobis D^2 statistics is defined as follows:

$$D^2 = b_1d_1+b_2d_2+.....+ b_pd_p = \sum_i \sum_j (\bar{X}_1 - \bar{X}_2)' W^{-1} (\bar{X}_1 - \bar{X}_2)$$

Where $\bar{x}_1^1$ is the mean value of 1st character in the first population and $\bar{x}_1^2$ is the mean of the 1st character in the second population. Here, the b_i values are to be estimated such that the ratio of variance between the populations to the variance within the population is maximized. In terms of variances and covariances, the D^2 distance between object 1 and object 2 can be obtained as;

$$D^2 = (\bar{X}_1 - \bar{X}_2)' W^{-1} (\bar{X}_1 - \bar{X}_2)$$

Where W^{-1} is the inverse of variance covariance matrix, $\bar{X}_1$ is the mean of first population, $\bar{X}_2$ is the mean of second population. It is used for quantitative data.

2.1.1.4 Chebychev's distance

This is a dissimilarity measure is based on the assumption that two objects are different if they differ in any one of the characteristics and is calculated as,

$$C_{ij} = Max \left|X_{ik} - X_{jk}\right|$$

2.1.2 Clustering techniques

These are the procedures used for clubbing together of similar objects into different clusters. There are different techniques for clustering *i.e*. Hierarchical technique and *K*-means clustering [13].

2.1.2.1 Hierarchical clustering

Hierarchical clustering assumes each of the '*n*' objects as individual clusters initially. Similar objects are combined together in successive fusions or dissimilar objects are divided in successive divisions in this clustering technique. There are different methods for linking the objects in different clusters. Single linkage (nearest neighbour) method, complete linkage (farthest neighbour) method and Tocher's method are common.

In single linkage, two individuals having minimum distance forms the first cluster. In next step, a third individual is joined with the initial cluster or another two nearer individuals are clustered together to form the second cluster. This is determined if the distance from the third individual to the first cluster is shorter than the distance between the two nearer individuals. Two objects having maximum distance between them constitute two groups. Next object either join one of the previous clusters or form its own cluster.

Tocher's method of clustering objects makes use of Mahalanobis D^2 statistic. D^2 values are arranged in ascending order and the two individuals having smallest distance between them is selected as the first cluster. Tocher suggested a cut off value which is equal to maximum among the minimum D^2 values. Addition of a third object is determined in such a way that average D^2 distance is less than the cut off value. Clustering continues until all the objects are included in one of the clusters [14].

2.1.2.2 K -means clustering

K-means clustering is a technique where we define some pre-specified number of clusters for grouping '*n*' objects. Only condition is that clusters are formed with minimum variance within clusters and maximum variance between clusters.

2.1.3 Cluster validity index

Davies-Bouldin index is one of the cluster validity measures to evaluate the clusters based on their compactness and separation from each other. Let $X_1, X_2, \ldots X_c$ are the clusters and Δx_i , $\delta (x_i, x_j)$ represents the intra cluster distance of i^{th} cluster and inter cluster distance between i^{th} and j^{th} cluster respectively. Then DB index is defined as,

$$DB = \frac{1}{c}\sum_{i=1}^{c} \max_{i \neq j} \frac{(\Delta x_i + \Delta x_j)}{\delta(x_i, x_j)}$$

DB index values should be less for good clustering *i.e*. small index values corresponds to clusters which are compact and well separated. Clustering algorithms that minimize the DB index values give optimum number of clusters.

3. RESULTS AND DISCUSSION

In order to classify the various panchayats based on the soil fertility status in Ernakulam district, cluster analysis was adopted [15]. Different clustering algorithms such as *K*-means clustering and hierarchical agglomerative clustering were adopted in the present study. Various dissimilarity measures such as Euclidean distance, squared Euclidean, Mahalanobis D^2 and Chebychev distance were estimated for clustering the panchayats. Various clustering procedures like single linkage, complete linkage and average linkage were followed under hierarchical clustering to generate clusters consisting of panchyats having similar soil fertility properties [16]. Tocher's clustering procedure was also employed to generate clusters using Mahalanobis D^2 distance measure. Comparison of various measures and clustering algorithms were completed using SPSS package. The study conducted by [17] was in accordance with the present study as different clustering techniques provided different number of clusters.

3.1 Hierarchical clustering of panchayats

3.1.1 Squared Euclidean distance

Squared Euclidean distance was selected for clustering the panchayats and the dissimilarity matrix was determined. Single linkage, complete linkage procedures were practiced along with average linkage method for squared Euclidean distance and the clustering of panchayats in each case is given in Table 1.

There were seven clusters under single linkage and average linkage clustering procedure and six clusters for complete linkage. Single linkage was another hierarchical clustering procedure and it did not provide proper clusters of panchayats in EKM district.

It is evident from the Table 1 that Ayyampuzha, Chengamanadu, Keerampara, Thirumaradi and Thuravur formed as single clusters when single linkage was adopted. Five panchayats out of 17 were in individual clusters. Only two clusters were having more than one panchayat in it and all other panchayats reamined as individual clusters.

The panchayats Ayyampuzha, Chengamanadu, and Thuravur remained as single clusters under complete linkage and average linkage methods and Kadungallor, Kalady and Puthenvelikkra were also in same cluster under these clustering procedures. There were six clusters under complete linkage and seven clusters under average linkage. Ayyampuzha, Chengamanadu and Thuravur stood as individual clusters,(Table 1).

Table 1. Clustering of panchayats in EKM based on single linkage, complete linkage and average linkage method (Squared Euclidean distance).

Cluster	Panchayats coming under each cluster		
	Single linkage	Complete linkage	Average linkage
I	Ayyampuzha	Ayyampuzha	Ayyampuzha
II	Kadungalloor, Kalady, Puthenvelikkara	Kadungalloor, Kalady, Puthenvelikkara	Kadungalloor, Kalady, Puthenvelikkara
III	Thuravur	Thuravur	Thuravur
IV	Chengamanadu	Chengamanadu	Chengamanadu
V	Keerampara,	Keerampara, Thirumaradi, Mudakkuzha,Nedumbassery, Piravom	Keerampara, Mudakkuzha, Thirumaradi
VI	Mudakkuzha, Kakkad, Vengola, Manjapra, Vengoor, Nedumbassery, Piravom, Pampakuda, Pothanikkad	Kakkad, Vengola, Manjapra, Vengoor, Pampakuda, Pothanikkad	Kakkad, Vengola, Manjapra, Vengoor
VII	Thirumaradi		Nedumbassery, Piravom, Pampakuda, Pothanikkad

3.1.2 Euclidean distance

Euclidean distance was also selected as the distance measure and single linkage, complete linkage and average linkage procedures were performed under this distance measure. Average linkage using Euclidean distance as the distance measure gave the same clustering pattern as that of squared Euclidean (Table 2). Unlike squared Euclidean, average linkage and complete linkage using Euclidean distance resulted in seven clusters and single linkage method had six clusters. Ayyampuzha, Chengamanadu, Keerampara and Thuravur were formed as clusters with single panchayat in it and hence intra cluster distance was zero. Except Keerampara which was a single cluster for single linkage method only, Ayyampuzha, Chengamanadu and Thuravur were also found to be single clusters when complete and average linkage methods were used. Similar findings were put forwarded by [18].

Cluster	Panchayats coming under each cluster

	Sinkle linkage	Complete linkage	Average linkage
I	Ayyampuzha	Ayyampuzha	Ayyampuzha
II	Thuravur	Thuravur	Thuravur
III	Chengamanadu	Chengamanadu	Chengamanadu
IV	Kadungalloor, Kalady, Puthenvelikkara	Kadungalloor, Kalady, Puthenvelikkara	Kadungalloor, Kalady, Puthenvelikkara
V	Keerampara	Keerampara, Mudakkuzha, Thirumaradi, Nedumbassery, Piravom	Keerampara, Mudakkuzha, Thirumaradi
VI	Nedumbassery, Piravom Kakkad, Mudakkuzha, Pampakuda, Pothanikkad, Vengola, Vengoor, Manjapra, Thirumaradi	Kakkad, Vengola, Manjapra, Vengoor, Pampakuda, Pothanikkad	Kakkad, Vengola, Manjapra, Vengoor
VII		Kadungalloor, Kalady, Puthenvelikkara	Nedumbassery, Piravom, Pampakuda, Pothanikkad,

Table 2.Clustering of panchayats in EKM based on single linkage, complete linkage and average linkage method (Euclidean distance).

Kadungalloor, Kalady and Puthenvelikkara came under a cluster when all these methods were practiced. This revealed that these three panchayats had almost similar soil characteristics. Similarly, Kakkad, Vengola, Manjapra and Vengoor were the panchayats which formed into a cluster and it was inferred that these panchayats also had similar soil properties.

Clusters obtained through Euclidean distance were similar as that of squared Euclidean distance. Ayyampuzha, Chengamanadu and Thuravur retained as individual clusters with intra cluster distance zero.

3.1.3 Mahalanobis D^2 distance

Panchayats coming under each cluster as per Mahalanobis D^2 distance is given in Table 3. This distance measure provided seven clusters when Tocher's clustering method was adopted. Manjapra and Keerampara were the panchayats which retained as single cluster with intra cluster distance zero when Mahalanobis D^2 was selected as the distance measure. This result is in line with the results of the research work done by [5] on clustering of rice genotypes.

Table 3. Clustering of panchayats in EKM based on Tocher's method

Cluster	Panchayaths coming under each cluster
I	Kakkad, Kalady, Vengoor
II	Kadungalloor, Pothanikkad, Puthenvelikkara, Pampakuda, Thirumaradi
III	Chengamanadu, Vengola
IV	Ayyampuzha, Mudakkuzha
V	Nedumbassery, Piravom
VI	Manjapra
VII	Keerampara

3.1.4 Chebychev distance

Clustering of panchayats based on average linkage, single linkage and complete linkage methods using Chebychev distance as the dissimilarity measure is given below in Table 4. There were 12 clusters into which panchayats were grouped when single linkage method was adopted. Other than Ayyampuzha, Chengamanadu and Thuravur which were separate clusters in all the clustering procedures, Manjapra, Vengoor, Thirumaradi, Puthenvelikkara, Mudakkuzha and Keerampara stood as individual clusters. Kadungalloor, Kalady and Puthenvelikkara were formed as a single cluster inferring that these three panchayats had almost similar soil nutrient status. Complete linkage and average linkage gave seven clusters. Mudakkuzha, Keerampara and Thirumaradi came under the same cluster when both average and complete linkage methods were adopted.

Table 4. Clustering of panchayats in EKM based on Chebychev distance under different methods

Cluster	Panchayats coming under each cluster		
	Single linkage	Complete linkage	Average linkage
I	Ayyampuzha	Ayyampuzha	Ayyampuzha
II	Thuravur	Thuravur	Thuravur
III	Kadungalloor, Kalady	Kadungalloor, Kalady, Puthenvelikkara	Kadungalloor, Kalady, Puthenvelikkara
IV	Chengamanadu	Chengamanadu, Nedumbassery, Piravom	Chengamanadu
V	Mudakkuzha	Mudakkuzha, Keerampara, Thirumaradi	Mudakkuzha, Keerampara, Thirumaradi
VI	Vengola, Kakkad,	Vengola, Kakkad, Manjapra, Vengoor	Vengola, Kakkad, Manjapra, Vengoor
VII	Pothanikkad, Pampakuda, Nedumbassery, Piravom	Pothanikkad, Pampakuda	Pothanikkad, Pampakuda Nedumbassery, Piravom
VIII	Vengoor		
IX	Puthenvelikkara		
X	Keerampara		
XI	Manjapra		
XII	Thirumaradi		

Chebychev distance also provided somewhat similar clustering patterns as that of squared Euclidean and Euclidean distance. Only difference was in the number of clusters formed when single linkage

method used under the three distance measures. Chebychev distance gave 12 clusters while Euclidean and squared Euclidean gave 6 and 7 clusters respectively.

3.2 *K*-means clustering

K –means clustering defines the number of clusters in advance. In the present study, number of clusters was taken as 5, 6 and 7 in *K*- means clustering. When *k* = 5, *K* = 6 and *K*= 7 panchayats of Ernakulam was classified into five, six and seven groups respectively and are presented in Table 5.

Only Ayyampuzha was as an individual cluster for *K*= 5, 6 and 7. Thuravur became a single cluster when *k* =6. Nedumbassery, Piravom and Chengamanadu came under cluster II for *K*= 5,6 and 7.Kadungalloor, Kalady and Puthevelikkara belonged to same cluster which suggested that they had similar soil properties. The same clustering procedure was adopted by [7] for partitioning locations based on soil properties.

Thuravur, which was in the same cluster with other panchayats, now formed as a new individual cluster when *K* =6 was selected. Panchayats were again redistributed among different clusters when *K*- means clustering with *K* =7 were performed. Now, Ayyampuzha, Keerampara and Thuravur were in different clusters with intra cluster distance zero *i.e*. those cluster had single panchayats as the entity.

Table 5. Clustering of panchayats in EKM based on K- means clustering

Cluster	Panchayats coming under each cluster		
	K =5	K =6	K =7
I	Ayyampuzha	Ayyampuzha	Ayyampuzha
II	Chengamanadu, Nedumbassery, Piravom	Chengamanadu, Nedumbassery, Piravom	Chengamanadu, Nedumbassery, Piravom
III	Kadungalloor, Kalady, Puthenvelikkara	Kadungalloor, Kalady, Puthenvelikkara	Kadungalloor, Kalady, Puthenvelikkara
IV	Kakkad, Manjapra, Pampakuda, Pothanikkad, Thuravur, Vengola, Vengoor	Kakkad, Manjapra, Pampakuda, Pothanikkad, Vengola, Vengoor	Kakkad, Manjapra, Pothanikkad, Vengola, Vengoor
V	Keerampara, Mudakkuzha, Thirumaradi	Keerampara, Mudakkuzha, Thirumaradi	Mudakkuzha, Pampakuda, Thirumaradi
VI		Thuravur	Thuravur
VII			Keerampara

It was concluded that Ayyampuzha, Thuravur and Chengamanadu were the panchayats with different soil characteristics and did not form any group with other panchayats based on the soil fertility

parameters. In most of the clustering procedures adopted, Kadungalloor, Kalady and Puthenvelikkara came under a cluster based on their soil properties which indicated that these three panchayats had almost similar soil properties.

3.3. Cluster validity index- panchayats of Ernakulam

Being unsupervised procedure, cluster analysis need evaluation of the results of different clustering procedures. Cluster validity means identifying the clusters that best fit to the given data. Davies-Bouldin index was one of such measures used for cluster validation which was used in the present study (Table 6). Comparatively low values are preferred for good clustering procedure in this method.

Table 6. Cluster validity index for different clustering procedures (EKM)

Distance measure	Linkage method	DB index score
Squared Euclidean	Average linkage	0.412
	Complete linkage	0.427
Euclidean distance	Average linkage	**0.306**
	Complete linkage	0.894
Chebychev distance	Average linkage	0.458
	Complete linkage	0.383
Mahalanobis D^2		**0.120**
K – means clustering	*K* =5	0.566
	K =6	0.467
	K =7	0.497

Single linkage method was not considered as most of the panchayat was retained as single clusters with intra cluster distance zero. A comparison between average linkage and complete linkage was carried out using the D-B index [19]. D- B index should be less for the optimum clustering pattern and here Mahalanobis D^2 was found to be the best clustering measure followed by average linkage method with Euclidean as the distance measure. Cluster means were determined for Mahalanobis D^2 and the results are given in Table 7.

Table 7. Cluster mean based on Mahalanobi's D^2 (EKM)

Cluster no.	pH	EC	OC	P	K	Ca	Mg	S	B	Fe	Mn	Zn	Cu
I	5.297	0.144	1.959	47.067	298.562	530.051	71.953	22.364	0.546	85.713	26.012	3.673	2.805
II	4.880	0.129	1.668	49.522	191.996	520.996	29.114	28.579	0.743	60.605	17.956	2.122	4.346
III	4.640	0.197	1.573	86.173	231.798	414.339	55.541	24.920	0.906	35.649	12.252	1.349	4.266
IV	5.207	0.151	1.911	56.747	346.930	522.571	108.995	13.525	0.717	76.649	29.685	2.436	3.313
V	5.083	0.103	1.394	55.613	165.148	369.000	80.500	20.948	0.863	38.035	13.610	1.797	1.505
VI	5.284	0.284	1.789	82.774	209.016	490.526	19.684	57.653	1.076	115.790	27.500	1.621	2.011
VII	5.171	0.066	1.928	31.203	139.112	274.655	41.172	21.247	1.015	158.241	21.764	5.709	5.222

Keerampara and Manjapra remained as separate panchayats when Mahalanobis D^2 was practiced. Even though both panchayats come under the agro ecological unit 3.1 (southern and central foot hills), they were not found together under a cluster as they would be having dissimilar soil characteristics [20]. Keerampara soils were strongly acidic with comparatively lower available Ca content (274 mg kg-1) and high Fe status (158 mg kg-1). Some parts of Keerampara had proximity with water bodies. Nedumbassery and Piravom belonged to cluster V with EC (0.10 dS m-1), OC (1.39 %), K (165.14 kg ha-1) and S (20.94 mg kg-1). Cluster IV comprised Ayyampuzha and Mudakkuzha panchayats with an average pH (5.2), EC (0.15), OC (1.91 %) and K (346.93 kg ha-1). Kadungalloor, Pothanikkad, Puthenvelikkara, Pampakuda and Thirumaradi clubbed together to form a cluster with 1.6 per cent OC, P (49.5 kg ha-1) and K (191.99 kg ha-1). Kakkad, Kalady and Vengoor came under cluster I as they had approximately the same EC (0.15-0.19 dS m-1), OC (2-2.4%) and Mg (71-73 mg kg-1). Chengamanadu which was reported as an individual cluster in all the clustering methods adopted, came along with Vengola under Tocher's method. OC ranged from 1.55 to 1.58 per cent, K (216.88-246.71 kg ha-1), Mg (54-56 mg kg-1) and B (0.89-0.91 mg kg-1) in these panchayats.

For all the clustering methods used except Mahalanobis D^2, Ayyampuzha retained as a single cluster. Hierarchical clustering also provided information that Ayyampuzha did not form any group with other panchayats. Ayyampuzha had moderately acidic soils (pH - 5.8) with high available K (570 kg ha-1) while Chengamanadu had high available P (113 kg ha-1) and low Ca content (300 mg kg-1). Ayyampuzha comes under the agro ecological unit 4.1 (Southern high hills) and lies near to waterbodies. Thuravur was deficient in available B (0.28mg kg-1). Kadungalloor, Kalady and Puthenvelikkara came under cluster IV and the soils in this cluster recorded with high K (342.82 kg ha-1) and Ca (619.82 mg kg-1). The cluster V comprised of Keerampara, Mudakkuzha and Thirumarady and the soils were strongly acidic with comparatively lower P (33.84 kg ha-1) status. The soils of panchayats in cluster VI were moderately acidic with sufficient quantity of soil nutrients. Cluster means calculated for different clusters based on Euclidean distance with average linkage method is presented in Table 8.

Table 8. Cluster mean based on Euclidean distance (EKM)

Cluster no.	pH	EC	OC	P	K	Ca	Mg	S	B	Fe	Mn	Zn	Cu
I	5.84	0.15	1.73	63.72	570.21	670.70	109.77	8.12	0.50	19.99	16.40	2.30	1.17
II	5.09	0.15	1.74	41.61	112.98	616.44	33.68	16.05	0.28	37.77	29.12	6.51	4.34
III	4.24	0.25	1.56	113.59	246.71	300.71	56.79	27.29	0.89	43.97	7.89	1.29	7.30
IV	4.99	0.17	1.72	71.44	342.01	619.82	41.81	25.06	0.64	85.70	15.21	1.78	5.92
V	4.68	0.14	2.16	33.84	102.87	336.00	60.32	28.49	0.99	134.73	24.89	3.13	4.75
VI	5.40	0.16	1.71	56.91	242.05	497.79	54.18	33.66	0.68	71.39	23.35	2.81	2.28
VII	5.22	0.15	1.52	58.70	217.07	470.01	57.12	26.77	0.80	70.42	21.28	2.57	2.02

Among the average and complete linkage methods under Euclidean and squared Euclidean distances, clustering based on average linkage method provided better clusters as the DB index was low. DB index of complete linkage under squared Euclidean was 0.427 which was higher than that of average linkage (0.412). Average linkage method had small DB index (0.306) under Euclidean distance compared to squared Euclidean.

4. CONCLUSION

Soils exhibit a high degree of heterogenity with respect to place and climate. Spatial distribution of soil nutrients also vary with locations [21]. Fertilizer recommendations are given based on the soil test results of the locations and if the locations having similar soil characteristics can be grouped based on soil fertility status, recommendations could be given easily [22].

Clustering of panchayats in Ernakulam were carried out using hierarchical clustering and *K*- means clustering. Different distance measures like euclidean distance, squared Euclidean, Chebychev distance and Mahalanobis D^2 along with different linkage methods – single, complete and average linkage were used for clustering the panchayats based on soil fertility status. . Davies- Bouldin index was determined for each clustering procedure and Mahalanobis D^2 was found to be the best clustering method followed by euclidean distance with average linkage method as the D-B index was small for Mahalanobis D^2.

Clustering of panchayats in Ernakulam based on Mahalanobis D^2 resulted in seven clusters. Keerampara and Manjapra remained as individual clusters. Keerampara soils were strongly acidic with EC value 0.06 dS m-1. Available Ca was observed to be comparatively lowest (274 mg kg-1) while Fe content was the highest (158 mg kg-1). Available Mg was very low in Manjapra (19 mg kg-1) compared to other panchayats. On the contrary, S content was more in this panchayat (57 mg kg-1). Nedumbassery and Piravom belonged to the same cluster due to the similarities in their soil characteristics. There were same EC (0.10 dS m-1), OC ranged from 1.2 to 1.5 per cent, K ranged from 161 to 168 kg ha-1 and S ranged from 20 to 21 mg kg-1 in these two panchayats. Kakkad, Kalady and Vengoor belonged to the same cluster as EC (0.15-0.19 dS m-1), OC (2-2.4%) and Mg (71-73 mg kg-1) had approximately same values.

ACKNOWLEDGEMENT

This research was supported by Kerala Agricultural University (KAU) and we thank our colleagues from KAU who provided insight and expertise that greatly assisted the research. We would also like to show our gratitude to all those who gave assistance and comments for completing this research.

REFERENCES

1. Das DK. Introductory Soil Science (Indian reprint, 2016).Kalyani Publishers, New Delhi; 1996.
2. Bernstein L. Effects of salinity and sodicity on plant growth. Annu. Rev. Phytopathol. 1975; 13:295-312.
3. Das MN, Giri NC. Design and analysis of experiments. 3rd ed. New age International Publishers, New Delhi; 1979.
4. Dawes L, Goonetilleke A. Using multivariate analysis to predict the behavior of soils under effluent irrigation. Water Air Soil Pollut. 2006; 172 (1-4): 109-127.
5. Bharadwaj C, Satyavathi CT, Subramanyam D. Evaluation of different classificatory analysis methods in some rice (*Oryza sativa*) collections. Indian J. Agric. Sci. 2001; 71(2): 123-125.
6. Temp M, Filzmoser P, Reimann C. Cluster analysis applied to regional geochemical data: Problems and possibilities. [on-line]. 2006. Available: https://www.sciencedirect.com/science/article/pii/S088329270800125X.
7. Altdorff D, Dietrich P. Cluster analysis of geophysical field data: an approach for reasonable partitioning of sites. 2010. In: 19th World Congress of Soil Science on 'Soil Solutions for a Changing World' 1 – 6 August 2010, Brisbane, Australia [On-line].
8. Ye R, Wright AL. Multivariate analysis of chemical and microbial properties in histosols as influenced by land-use types. Soil Till. Res. 2010; 100:94-100.
9. Morrison JM, Goldhaber MB, Ellefsen KJ, Mills CT. Cluster analysis of a regional-scale soil geochemical dataset in northern California. J. Int. Assoc. Geochem. 2011; 26: S105-S107.
10. Legany C, Juhasz S, Babos B.. Cluster validity measurement techniques. Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006; 388-393.
11. Ahmed F, Fakhruddin ANM, Imam MDT, Khan N, Khan TA, Rahman MM, *et al*. Spatial distribution and source identification of heavy metal pollution in roadside surface soil: a study of Dhaka Aricha highway, Bangladesh. Ecol. Process. 2016; 5(2): 1-16.
12. Bansod BS, Pandey OP, Rajesh NL. Analysis and delineation of spatial variability using geo-sensed apparent electrical conductivity and clustering techniques. Int. J. Agric. Biol. 2012; 14: 481–491.
13. Granato D, Santosa JS, Eschera GB, Ferreira BL, Maggio RM. Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. Trends Food Sci. Technol. 2018; 72: 83–90.
14. Mahalanobis PC. On the generalized distance in statistics. Proceedings of the National Institute of sciences (Culcutta), 1936; 2: 49-55.
15. de Araujoneto JR, Gomes FEF, de Q Palacio HA, da Silva EB, Brasil PP. Similarity of soils with regard to salinity in the perennial valley of trussu river, Ceará. Irriga, Botucatu. 2016; 21(2): 327-341.
16. Ellefsen KJ, Smith DB. Manual hierarchial clustering of regional geochemical data using a Bayesian finite mixture model. J. Int. Assoc. Geochem. 2016; 75: 200-210.
17. Guler C, Thyne GD, McCray JE, Turner AK. Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeol. J. 2002; 10:455–474.
18. Bakry AB, Ibrahim OM, Abd El-Fattah Elewa T, El-Karamany MF. Performance assessment of some flax (Linum usitatissimum l.) varieties using cluster analysis under sandy soil conditions. Agric. Sci. 2014; 5: 677-686.
19. Ansari Z, Babu AV, Azeem MF, Ahmed W. Quantitative evaluation of performance and validity indices for clustering the web navigational sessions. World Comput. Sci. Inf. Technol. J. 2011; 1(5): 217-226.

20. GoK (Government of Kerala). Fertility of Soils of Kerala. State planning board, Thiruvananthapuram; 2013.
21. Al-Farhud A, Al-Sewailem M, Usman ARA. Status of selenium and trace elements in some arid soils cultivated with forage plants: A case study from Saudi Arabia. Int. J. Agric. Biol. 2017; 19: 85–92.
22. KAU (Kerala Agricultural University). Package of Practices Recommendations: Crops (15th Ed.). Kerala Agricultural University, Thrissur; 2016.