International Journal of Scientific Research and Engineering Development-– Volume 7 Issue 6, Nov-Dec 2024
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Analysis of Soil Parameter Variability in Vadodara District Using Clustering Techniques

Bhagirath Prajapati*, Priyanka Puvar**
*(Computer Engineering, CVM University, V. V. Nagar
Email: bhagirath@adit.ac.in)
** (Computer Engineering, CVM University, V. V. Nagar
Email: priyanka.puvar@cvmu.edy.in)

----------**********----------

Abstract:
A standalone system which will help the agriculture department to cluster the villages based on the soil's major parameters. In this project Clustering technique of Data Mining is going to be used. "Clustering is the process of organizing objects into groups whose members are similar in some way". Cluster analysis is a general term for a family of statistical classification methods that group objects. It comes under unsupervised learning category of Machine Learning. Machine Learning is a field of study that gives computers the ability to learn without being explicitly programmed. The idea is statistically to minimize within-group variability while maximizing among-group variability in order to produce relatively homogeneous groups that are distinct from one another. This project is to study the pattern of variability of major soil parameters simultaneously across the villages and to group the villages having the same pattern by considering data of villages. It will help the farmers to identify the amount and type of fertilizers to be used in their soil on the basis of soil parameters in their region. This information will be easily available to the farmer from the agriculture department of that region.

***Keywords* —** soil parameter, clustering, k-means, R language.

----------**********----------

I. INTRODUCTION

A general problem facing researchers in many areas is how to organize large observed data into meaningful structures. Many statistical techniques have been used to analyze such huge data. Agricultural Department is one such voluminous wherein there is scope to apply suitable statistical techniques to bring out possible variability pattern.

Cluster analysis is a general term for a family of statistical classification methods that group objects. The idea is statistically to minimize within-group variability while maximizing among-group variability in order to produce relatively homogeneous groups that are distinct from one another. Cluster analysis can be performed by using two groups of techniques; hierarchical and non-hierarchical. However, the most commonly applied technique is hierarchical.

II. PREVIOUS WORK

Soil health is a basic requirement for crop production and Soil Testing Laboratories (STLs) were established decades earlier. STLs routinely collected soil samples in the order of one lakh per year and after every five years; soil fertility maps were prepared on the basis of five lakh data. But such maps were, though useful to the administrators, of no use to the specific farmer. With the introduction of SHC (Soil Health Card), the individual farmer is benefited. The SHC contains information such as farmer's name, account number, survey number, soil fertility, soil Organic carbon, available P & K, pH and EC values and their Low, Medium or High status, general fertilizer dose recommended and soil test based fertilizer and manure rates to be given for each crop growth by the farmer. Until now various statistical techniques have been applied to cluster soil parameters. This study aims at clustering the soil parameters using data mining approach through R language.[2]

III. PROPOSED WORK

The aim of this study is to cluster the data obtained from the Soil Health Card Scheme proposed by The Government Of India. It mainly focuses on four parameters of the soil, i.e. PH value, carbon, phosphorus, and potash. The results obtained from cluster analysis will help us to take a decision for a particular region, i.e. we will be able to know about the deficiency of a particular soil parameter in that region.

IV. METHODS AND ALGORITHMS

A. R Language and R studio

R is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.[10] R is an integrated suite of software facilities for data manipulation, calculation, and graphical display. It includes an effective data handling and storage facility, a suite of operators for calculations on arrays, in particular matrices, a large, coherent, integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and display either on-screen or on hard copy, and a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

R can be extended (easily) via packages. There are several packages supplied with the R distribution and much more are available through the CRAN family of Internet sites covering a very wide range of modern statistics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. [5]

R Studio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. R Studio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) or in a browser connected to R Studio Server or R Studio Server Pro (Debian/Ubuntu, RedHat / CentOS, and SUSE Linux).

R is a programming language for statistical computing and graphics. R Studio is available in two editions: R Studio Desktop, where the program is run locally as a regular desktop application; and R Studio Server, which allows accessing R Studio using a web browser while it is running on a remote Linux server. Prepackaged distributions of R Studio Desktop are available for Windows, OS X, and Linux.

B. K-means Clustering Algorithm

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes the different result. So, the better choice is to place them as much as possible far away from each other.[4]

The most common partitioning method is the K-means cluster analysis. Conceptually, the K-means algorithm:

1. Selects K centroids (K rows chosen at random)
2. Assigns each data point to its closest centroid
3. Recalculates the centroids as the average of all data points in a cluster (i.e., the centroids are p-length mean vectors, where p is the number of variables)
4. Assigns data points to their closest centroids
5. Continues steps 3 and 4 until the observations are not reassigned or the maximum number of iterations (R uses 10 as a default) is reached.

Usage:

kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"), trace=FALSE)

x: a numeric matrix of data, or an object that can be coerced to such a matrix

centers: either the number of clusters, say k, or a set of initial (distinct) cluster centers.

iter.max: the maximum number of iterations allowed.

nstart: if centers is a number, how many random sets should be chosen?

algorithm: character: may be abbreviated. Note that "Lloyd" and "Forgy" are alternative names for one algorithm.

object: an R object of class "kmeans", typically the result ob of ob<- kmeans(..).

trace: logical or integer number, currently only used in the default method ("Hartigan-Wong"): if positive (or true), tracking information on the progress of the algorithm is produced. Higher values may produce more tracing information.

C. Normalization of Data

For clustering any dataset using kmeans algorithm the data should be in a normalized form. In R studio there is one library called clusterSim using which we can normalize our dataset into one common data format. In-library clusterSim there is data.Normalization is a function which is used for types of variable normalization formulas. [9]

Usage

data.Normalization (x,type="n0")14

x: vector, matrix or dataset

type: type of normalization; n0 to n11

D. Determining number of cluster

For clustering any dataset the main thing is how many numbers of a cluster should be for a particular dataset. To resolve this problem in R studio package NbClust can be used to determine the number of clusters.

Description

NbClust package for determining the best number of clusters. NbClust package provides 30 indices for determining the number of clusters and proposes to a user the best clustering scheme from the different results obtained by varying all combinations of a number of clusters, distance measures, and clustering methods. [6]

Usage

NbClust (data = NULL, diss = NULL, distance = "euclidean", min.nc = 2, max.nc = 15, method = NULL, index = "all", alphaBeale = 0.1)

data: matrix or dataset.
diss: dissimilarity matrix to be used. By default, diss=NULL.
distance: the distance measure to be used to compute the dissimilarity matrix. By default, distance="euclidean".
min.nc: minimal number of clusters, between 1 and (number of objects - 1)
max.nc: maximal number of clusters, between 2 and (number of objects - 1)
method: the cluster analysis method to be used.
index: the index to be calculated.
alphaBeale: significance value for Beale's index.

E. Data Visualization

After clustering any dataset the resulting number for cluster and representation of data must proper. For that, there are many data visualization packages in R studio.(Fig.1,2)

For example

plot(, col =(ob$cluster +1) , main="K-Means result with 2 clusters", pch=20, cex=2)[5]

V. DATASETS

In the dataset, we have taken the 20 samples from each village of 12taluka of Vadodara district.

Sr. No.	Taluka Name	No. of Villages
1	Chhota Udaipur	42
2	Dabhoi	29
3	Jetpurpavi	54
4	Karajan	17
5	Kavvat	30
6	Nasvadi	41
7	Padra	20
8	Sankheda	47
9	Savli	33
10	Sinor	10
11	Vadodara	11
12	Vaghodia	25

VI. RESULT

Sr. No.	Taluka Name	No. of Clusters
1	Chhota Udaipur	5
2	Dabhoi	2
3	Jetpurpavi	3
4	Karajan	2
5	Kavvat	6
6	Nasvadi	2
7	Padra	6
8	Sankheda	2
9	Savli	3
10	Sinor	8
11	Vadodara	3
12	Vaghodia	4

VII. ANALYSIS

1. **Distribution of Chhota Udaipur Villages in clusters.**

Maximum numbers of villages i.e., 17 were observed in cluster 5. (Table 1)

Among the mean values of phosphorus for different clusters, the value (61.9) of the forth cluster was minimum whereas the value (90) of the second cluster was of maximum.

Among the mean values of potass for different clusters, the value (215) of the second cluster was minimum whereas the value (409.88) of the forth cluster was of maximum.

2. **Distribution of Dabhoi Villages in clusters.**

Maximum numbers of villages i.e., 18 were observed in cluster 1. (Table 2)

Among the mean values of potass for different clusters, the value (225.7) of the first cluster was minimum whereas the value (233.71) of the second cluster was of maximum.

3. **Distribution of JetpurPavi Villages in clusters.**

Maximum numbers of villages i.e., 25 were observed in cluster 1. (Table 3)

Among the mean values of pH for different clusters, the value (6.50) of the second cluster was minimum whereas the value (6.94) of the first cluster was of maximum.

4. **Distribution of Karjan Villages in clusters**

Maximum numbers of villages i.e., 9 were observed in cluster 2. (Table 4)

Among the mean values of carbon for different clusters, the value (0.54) of the second cluster was minimum whereas the value (4.009) of the first cluster was of maximum.

Among the mean values of potass for different clusters, the value (212.42) of the first cluster was minimum whereas the value (372.39) of the second cluster was of maximum.

5. **Distribution of Kavvat Villages in clusters.**

Maximum numbers of villages i.e., 9 were observed in cluster 5. (Table 5)

Among the mean values of phosphorus for different clusters, the value (27.58) of the second cluster was minimum whereas the value (84.67) of the forth cluster was of maximum.

Among the mean values of potass for different clusters, the value (335.05) of the first cluster was minimum whereas the value (566.68) of the forth cluster was of maximum.

6. **Distribution of Nasvadi Villages in clusters.**

Maximum numbers of villages i.e., 21 were observed in cluster 1. (Table 6)

Among the mean values of potass for different clusters, the value (301.34) of the second cluster was minimum whereas the value (351.18) of the first cluster was of maximum.

7. **Distribution of Padra Villages in clusters.**

Maximum numbers of villages i.e., 6 were observed in cluster 5. (Table 7)

Among the mean values of potass for different clusters, the value (187.4) of the third cluster was minimum whereas the value (815.25) of the second cluster was of maximum.

Among the mean values of carbon for different clusters, the value (0.294) of the sixth cluster was minimum whereas the value (0.858) of the second cluster was of maximum.

8. **Distribution of Sankheda Villages in clusters.**

Maximum numbers of villages i.e., 36 were observed in cluster 2. (Table 8)

Among the mean values of potass for different clusters, the value (232.68) of the second cluster was minimum whereas the value (348.96) of the first cluster was of maximum.

Among the mean values of phosphorus for different clusters, the value (45.21) of the second cluster was minimum whereas the value (80.01) of the first cluster was of maximum.

9. **Distribution of Savli Villages in clusters.**

Maximum numbers of villages i.e., 16 were observed in cluster 1. (Table 9)

Among the mean values of potass for different clusters, the value (242.02) of the second cluster was minimum whereas the value (380.37) of the third cluster was of maximum.

Among the mean values of phosphorus for different clusters, the value (36.36) of the third cluster was minimum whereas the value (64.21) of the first cluster was of maximum.

10. **Distribution of Sinor Villages in clusters.**

Maximum numbers of villages i.e., 2 were observed in cluster 5, 6. (Table 10)

Among the mean values of potass for different clusters, the value (342.3) of the second cluster was minimum whereas the value (402.75) of the third cluster was of maximum.

Among the mean values of phosphorus for different clusters, the value (40.35) of the forth cluster was minimum whereas the value (81.05) of the eighth cluster was of maximum.

11. **Distribution of Vadodara Villages in clusters.**

Maximum numbers of villages i.e., 4 were observed in cluster 1, 2. (Table 11)

Among the mean values of potass for different clusters, the value (389.81) of the second cluster was minimum whereas the value (602) of the third cluster was of maximum.

12. **Distribution of Vaghodia Villages in clusters.**

Maximum number of villages i.e., 9 was observed in cluster 4. (Table 12)

Among the mean values of potass for different clusters, the value (315.06) of the first cluster was minimum whereas the value (393.05) of the forth cluster was of maximum.

VIII. CONCLUSION

The information about the soil parameters of individual farmer is obtained from the Soil Health Card Scheme of the government but if a decision needs to be taken on a larger scale for a particular region specifically for business purpose then Cluster Analysis should be done on the soil parameters. Through the results obtained from cluster analysis various decisions can be taken.

ACKNOWLEDGMENT

It gives us immense pleasure to express our deepest gratitude to Prof. Bhagirath Prajapati, Associate Professor, Department Of Computer Engineering, A.D. Patel Institute Of Technology and Prof. Priyanka Puvar, Assistant Professor, Department Of Computer Engineering , A.D. Patel Institute Of Technology for their keen interest, constant encouragement and inspiration throughout the course of this study.

REFERENCES

[1] Marcos M. Campos; Boriana L. Milenova; Mark A. McCracken; "ENHANCED K-MEANS CLUSTERING"

[2] Abbasali N. Khokhar; "CLUSTERING OF VILLAGES BASED ON SOIL PARAMETERS – A CASE STUDY OF PANCHMAHAL DISTRICT"

[3] Jeysenthil.KMS; Manikandan.T; Murali.E; "THIRD GENERATION AGRICULTURAL SUPPORT SYSTEM DEVELOPMENT USING DATA MINING"

[4] Andrew Moore: "KMEANS AND HIERARCHICAL CLUSTERING TUTORIAlSLIDE" http://www2.cs.cmu.edu/~awm/tutorials/kmeans.html

[5] Christopher Gandrud; "REPRODUCIBLE RESEARCH WITH R AND RSTUDIO SECOND EDITION"

[6] Malika Charrad;Nadia Ghazzali;Veronique Boiteau;Azam Niknafs; "NBCLUST: AN R PACKAGE FOR DETERMINING THE RELEVANT NUMBER OF CLUSTERS IN A DATA SET"

[7] Brian T. Luke: "KMeans Clustering" http://fconyx.ncifcrf.gov/~lukeb/kmeans.html

[8] Ashok Kumar. D,Kannathasan. N, "A Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining"

[9] Walesiak M, Dudek A (2014). clusterSim: Searching for Optimal Clustering Procedure for a Data Set. R package version 0.43-4, URL http://CRAN.R-project.org/package= clusterSim.

[10] R Core Team (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

TABLE I. ISTRIBUTION OF CHHOTA UDAIPUR VILLAGES IN CLUSTERS.

Cluster	Village ID	Mean Value			
		PH	CARBON	PHOSPHORUS	POTASS
1	V11569, V11575, V11576, V11621, V11636, V11639, V11659, V11717.	7.48	0.43	74.21	296.51
2	V11619	7.10	0.58	90.00	215.00
3	V11577, V11578, V11645, V11651, V11653, V11660, V11716, V48810, V11603.	7.38	0.42	78.54	375.43
4	V11565, V11568, V11572, V11574, V11584, V11609, V48066	7.45	0.52	61.90	409.89
5	V11588, V11600, V11602, V11606, V11610, V11613, V11614, V11615, V11616, V11638, V11646, V11657, V11664, V11669, V11723, V11739, V11741.	7.49	0.54	72.80	337.97

TABLE II. DISTRIBUTION OF DABHOI VILLAGES IN CLUSTERS.

Cluster	Village ID	Mean Value			
		PH	CARBON	PHOSPHORUS	POTASS
1	V12490, V12513, V12519, V12528, V12531, V12535, V12561, V12562, V12565, V12588, V12589, V12594, V12595, V12597, V12604, V44823, V44828, V12545.	6.94	0.54	48.80	233.71
2	V12494, V12506, V12518, V12525, V12537, V12541, V12563, V12586, V12592, V44825, V44833.	7.13	0.56	42.07	225.70

TABLE III. DISTRIBUTION OF JETPUR PAVI VILLAGES IN CLUSTERS.

Cluster	Village ID	Mean Value			
		PH	CARBON	PHOSPHORUS	POTASS
1	V11750, V11751, V11759, V11774, V11778, V11789, V11791, V11792, V11808, V11812, V11823, V11834, V11835, V11842, V11850, V11851 ,V11867, V11879, V11901, V11905, V11909, V11917, V11920, V11921, V11943.	6.95	0.62	33.06	264.10
2	V11752, V11776, V11780, V11807, V11839, V11875, V11902, V11910, V11911, V11935, V11938, V11951, V50841.	6.50	0.60	32.46	245.44
3	V11745, V11796, V11811, V11813, V11824, V11825, V11857, V11877, V11899, V11900, V11907, V11916, V11924, V11950, V11809,V11908	6.83	0.53	34.98	247.72

TABLE IV. DISTRIBUTION OF KARJAN VILLAGES IN CLUSTERS.

Cluster	Village ID	Mean Value			
		PH	CARBON	PHOSPHORUS	POTASS
1	V13061, V13098, V13099, V13103, V13104, V13115, V13128, V19557,	7.44	4.01	32.59	212.43
2	V13040, V13042, V13044, V13045, V13052, V13060, V13064, V13116, V13123	7.47	0.55	49.26	372.39

TABLE V. DISTRIBUTION OF KAVVAT VILLAGES IN CLUSTERS.

Cluster	Village ID	Mean Value			
		PH	CARBON	PHOSPHORUS	POTASS
1	V12205, V12177, V12293	7.45	0.56	27.71	335.05
2	V12175, V12186, V12194, V12211, V12279	7.31	0.45	27.59	336.75
3	V12202, V12219, V12258, V12300	7.35	0.75	55.61	384.88
4	V12200, V12245, V12254, V12301	7.34	0.56	84.68	566.69
5	V12224, V12270, V12193, V12241, V12244, V12257, V12262, V12263, V12271	7.34	0.55	83.33	340.02
6	V12234, V12259, V12180, V12203, V12255	7.48	0.56	84.14	367.03

TABLE VI. DISTRIBUTION OF NASVADI VILLAGES IN CLUSTERS.

Cluster	Village ID	Mean Value			
		PH	CARBON	PHOSPHORUS	POTASS
1	V11960, V11961, V11970, V11985, V11989, V11990, V12000, V12007, V12011, V12018, V12022, V12023, V12028, V12029, V12042, V12054, V12058, V12070, V12089, V12106,	7.32	0.83	48.06	351.18

	V48024				
2	V11971, V11978, V12003, V12008, V12016, V12017, V12030, V12031, V12039, V12043, V12047, V12049, V12061, V12073, V12075, V12088, V12093, V12102, V12107, V12144	7.37	0.75	48.26	301.34

TABLE VII. DISTRIBUTION OF PADRA VILLAGES IN CLUSTERS.

Cluster	Village ID	Mean Value			
		PH	CARBON	PHOSPHORUS	POTASS
1	V12974, V13027	8.10	0.58	75.79	330.20
2	V12971	7.23	0.86	54.44	815.25
3	V12966, V12999	7.56	0.73	41.99	187.40
4	V12962, V12988, V13032, V13034	7.99	0.48	56.30	673.49
5	V12956, V12991, V13009, V13010, V13033, V13037	7.97	0.69	53.99	348.63
6	V12968, V12980, V12998, V13031, V13036	7.67	0.29	55.55	391.14

TABLE VIII. DISTRIBUTION OF SANKHEDA VILLAGES IN CLUSTERS.

Cluster	Village ID	Mean Value			
		PH	CARBON	PHOSPHORUS	POTASS
1	V12305, V12333, V12334, V12411, V12417, V12418, V12421, V12432, V12473, V12479, V12309	7.49	0.47	80.10	348.96
2	V12303, V12310, V12312, V12318, V12327, V12332, V12339, V12340, V12343, V12345, V12346, V12351, V12353, V12356, V12357, V12360, V12362, V12365, V12376, V12378, V12380, V12389, V12400, V12403, V12416, V12422, V12436, V12438, V12441, V12447, V12450, V12451, V12457, V12458, V12464, V12465	7.01	0.57	45.21	232.68

TABLE IX. DISTRIBUTION OF SAVLI VILLAGES IN CLUSTERS.

Cluster	Village ID	Mean Value			
		PH	CARBON	PHOSPHORUS	POTASS
1	V12618, V12632, V12633, V12637, V12666, V12667, V12673, V12689, V12690, V12691, V12714, V12730, V12731, V44841, V44842, V50833	7.64	0.55	64.22	281.20
2	V12694, V12695, V12699, V12705, V12708, V12709, V44847	7.13	0.74	40.17	242.03
3	V12668, V12675, V12684, V12696, V12697, V12710, V12711, V12734,	7.35	0.54	36.37	380.38

	V12739, V44843				

TABLE X. DISTRIBUTION OF SINOR VILLAGES IN CLUSTERS.

Cluster	Village ID	Mean Value			
		PH	CARBON	PHOSPHORUS	POTASS
1	V13155	7.49	0.47	65.60	368.50
2	V13136	7.38	0.65	65.20	342.30
3	V13139	7.60	0.49	69.85	402.75
4	V13156	7.26	0.66	40.35	345.70
5	V13140, V13145	7.34	0.60	72.35	373.43
6	V13162, V13163	7.53	0.68	44.83	356.25
7	V13143	7.53	0.55	68.90	361.25
8	V13147	7.44	0.57	81.05	362.75

TABLE XI. DISTRIBUTION OF VADODARA VILLAGES IN CLUSTERS.

Cluster	Village ID	Mean Value			
		PH	CARBON	PHOSPHORUS	POTASS
1	V12851, V12852, V12897, V12906	7.32	0.75	50.50	543.75
2	V12863, V12920, V12926, V12941	7.32	0.83	51.91	389.81
3	V12873, V12934, V12936	7.29	0.81	43.87	602.00

TABLE XII. DISTRIBUTION OF VAGHODIA VILLAGES IN CLUSTERS.

Cluster	Village ID	Mean Value			
		PH	CARBON	PHOSPHORUS	POTASS
1	V12809, V12813, V12824, V12834	7.12	0.47	40.70	315.06
2	V12752, V12833, V12835, V12837	7.54	0.41	46.70	347.88
3	V12758, V12766, V12782, V12783, V12796, V12799, V12825, V12836	7.34	0.79	45.51	331.25
4	V12762, V12775, V12780, V12816, V12817, V12819, V12822, V12823, V12818	7.38	0.49	31.79	393.06

FIGURE I GRAPHICAL PLOT REPRESENTATION OF ALL PARAMETER WITHIN 6 CLUSTER OF KAVVAT TALUKA

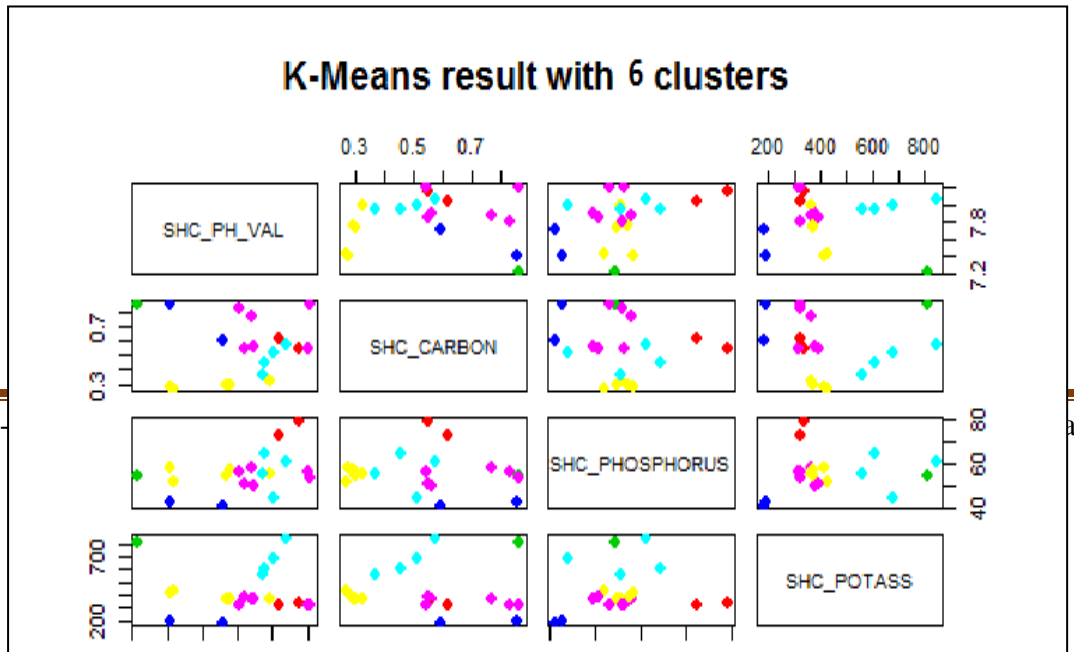

FIGURE II GRAPHICAL PLOT REPRESENTATION OF 2 PARAMETER WITHIN 6 CLUSTER OF KAVVAT TALUKA