IOSR Journal Of Applied Geology And Geophysics (IOSR-JAGG)
E-ISSN: 2321–0990, P-ISSN: 2321–0982.Volume 12, Issue 6 Ser. I (Nov. – Dec. 2024), 01-05
Www.Iosrjournals.Org

Data Visualisation To Understand How Data Is Structured Using K-Means And Hierarchical Cluster Analyses With Interactive Graphics

Amitabha Roy
Ex-Senior Director, Geological Survey Of India

Abstract
This study examines the structure of geothermal geochemisty data in India using k-means and hierarchical cluster analyses. The first attempt to list hot springs in India was made by Schlagintweit in 1852. The Geological Survey of India published a special publication titled 'Geothermal Atlas of India', and the government of India constituted a 'Hot Spring Committee' to examine the possibility of developing geothermal plants for power generation and other uses. In the Puga valley and Parvati projects, it is estimated that it is possible to harness 5000 MWh of geothermal energy from Puga valley, sufficient to sustain a 20 MWe power plant.
The GSI, the repository of most information concerning geological and related data in the country, included an R&D item No. 7/WB-5 for the development of a computerised system of geothermal database system referred to as GTHERMIS in its field season 1993-94 program. The computational strategy involves assigning each data point to the cluster with the nearest center (or "centroid"), recalculating the centroids after each assignment, and repeating the procedure until the clusters are no longer statistically different.
The k-means method, developed by MacQueen (1967), is one of the most widely used non-hierarchical methods, particularly suitable for large amounts of data. The data is scaled before cluster analysis, and the results are visualized using interactive graphics.
Keywords: *Hot springs, India, k-mean and hierarchial cluster, PCA, factor scores, Peninsula, extra-Peninsula, geothermal geochemistry*

Date of Submission: 15-11-2024 Date of Acceptance: 25-11-2024

I. Introduction

There are about 340 hot springs spread over different parts of India covering the Peninsular and Extra-Peninsular regions. The first attempt to list the hot springs in India was made by Schlagintweit in 1852. The Geological Survey of India has published a special publication titled 'Geothermal Atlas of India' (Ravi Shankar et. al.,1991), and the government of India constituted a 'Hot Spring Committee' to examine the possibility of developing geothermal plants for power generation and other uses. In the Puga valley and Parvati projects for utilisation of available geothermal resources for power generation. It is estimated to be possible to harness 5000 MWh of geothermal energy from Puga valley which is sufficient to sustain a 20 MWe power plant (Jonathan Craig et al., 2013). The GSI being the repository of most of the information concerning geological and other related data in the country, included in its field season 1993-94 program an R&D item No. 7/WB-5 for development of a computerised system of geothermal database system referred to as GTHERMIS (A.Roy,1994).

Computational strategy

In this study, geothermal geochemisty data is examined to discover underlying similarity patterns in data using the k-means cluster and hierarchical cluster analysis. The algorithm assigns each data point to the cluster with the nearest centre (or "centroid"). The centroids are recalculated after each assignment, and the procedure is repeated until the clusters are no longer statistically different. This aids in identifying patterns or structures in data.

k-means clustering

Algorithm aims at minimizing an objective function known as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

where,
'||x_i - v_j||' is the Euclidean distance between x_i and v_j.
'c_i' is the number of data points in i^{th} cluster.
'c' is the number of cluster centers.

Hierarchial clustering
D(r,s) = Trs / (Nr * Ns)

Where T_{rs} is the sum of all pairwise distances between cluster r and cluster s. N_r and N_s are the sizes of the clusters r and s, respectively. At each stage of hierarchical clustering, the clusters r and s , for which D(r,s) is the minimum, are merged.

Different linking methods and distances can be used to calculate the hierarchical cluster analysis. The linkage methods and distances are available for selection:

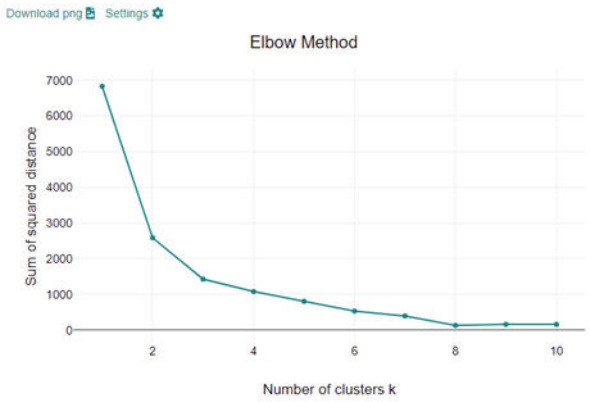
- **Linkage methods**: Single-linkage, Complete-linkage, Average-linkage,
- **Distances**: Euclidean, Manhattan, Maximum

Compared to the k-means cluster analysis, the number of clusters does not have to be specified in advance for the hierarchical cluster analysis. be specified in advance.
In the present study complete linkage and Euclidean distance approach are explored.

D(r,s) = Max { d(i,j) : Where object *i* is in cluster *r* and object *j* is cluster *s* }

The distance between every possible object pair (i,j) is computed, where object i is in cluster r and object j is in cluster s and the maximum value of these distances is said to be the distance between clusters r and s. The distance between two clusters is given by the value of the longest link between the clusters. At each stage of hierarchical clustering, the clusters r and s , for which D(r,s) is maximum, are merged.

In contrast to k-means cluster analysis, there is no need to specify the number of clusters beforehand.



k-means clustering

The k-Means method, which was developed by MacQueen (1967), is one of the most widely used non-hierarchical methods. It is a partitioning method, which is particularly suitable for large amounts of data.

Scaling data for k-means clustering: Since the variables under consideration do not have the same unit, the data has been first scaled before cluster analysis

The k-Means method, developed by MacQueen (1967), is one of the most widely used non-hierarchical partitioned methods. K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. First, an initial partition with k clusters (given number of clusters) is created. Then, starting with the first object in the first cluster, Euclidean distances of all objects to all cluster foci are calculated. These steps are repeated until each object is located in a cluster with the smallest distance to its centroid (center of the cluster). When you want to calculate a cluster analysis, often the big question is how many clusters should I take. The optimal cluster number is determined by elbow Curve method.In the present study 3 clusters are taken based on Elbow curve.

Hierarchical cluster analysis

The process of creating the dendrogram starts by computing a distance matrix between all pairs of objects. This distance matrix is then used to create a linkage matrix, which contains information about the distance between clusters at each stage of the analysis. The linkage matrix is then used to create the dendrogram, which shows how the clusters are related to each other.

Geothermal overlay on google map of India showing location of Hot Springs

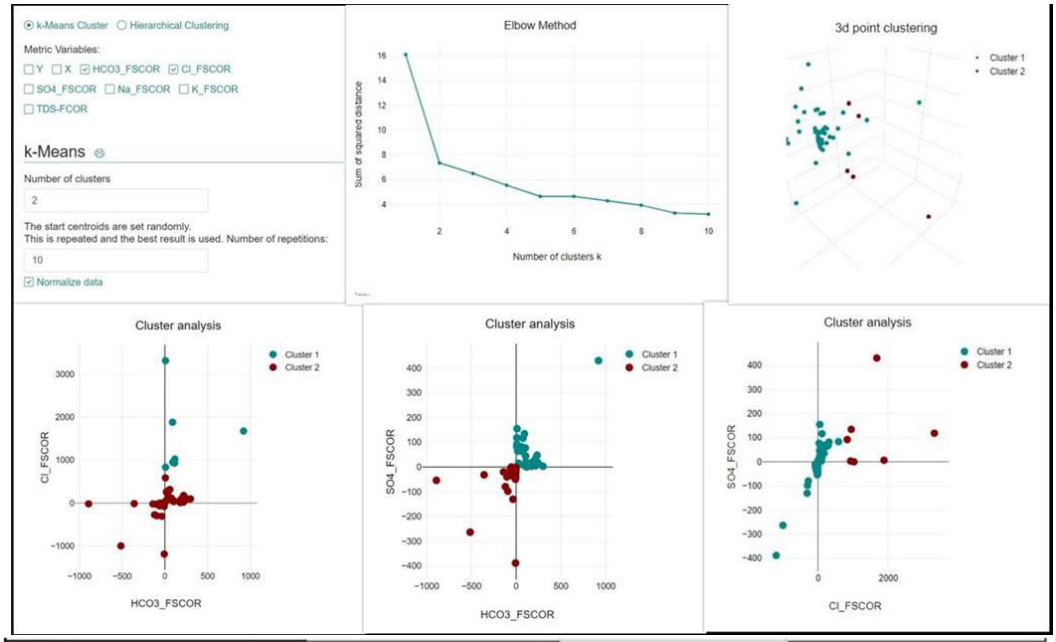
Geothermal Geochemistry Data

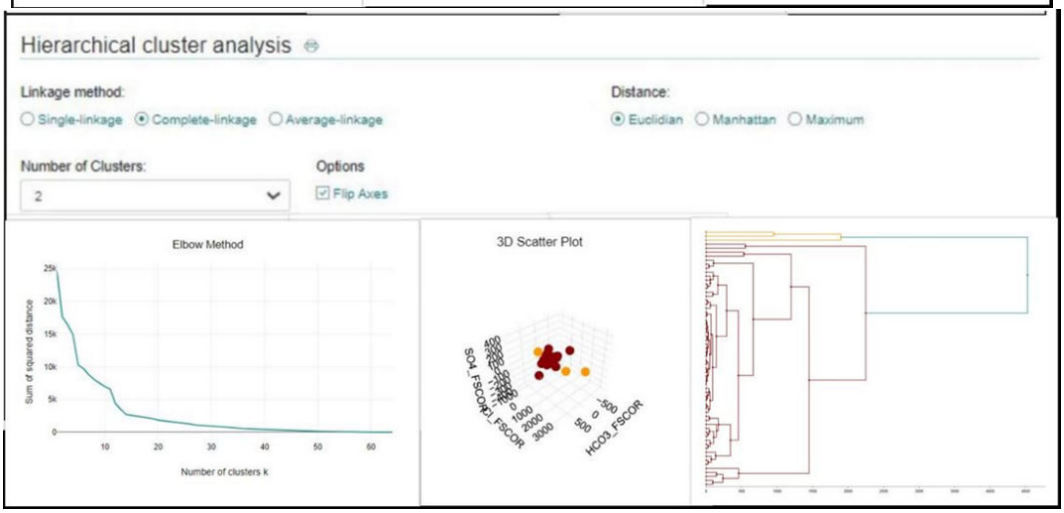
PCA Factor Scores						Original Raw Data				
HCO3_SCOR	Cl_FSCOR	SO4_FSCOR	Na_FSCOR	K_FSCOR		HCO3 mg/L	Cl mg/L	SO4 mg/L	Na mg/L	Na mg/L
92.4	948.75	134.4	396	16.02		300	163	62	210	13
203.4	113.85	15.36	66	5.34		170	133	36	88	19
189	89.7	21.12	42	22.25		490	855	1244	600	109
234	134.55	48	126	4.45		210	102	83	110	19
300	96.6	3.2	78	1.78		342	232	26	260	16
174	34.5	3.2	18	0.89		303	200	340	260	45
114	929.43	3.2	4086	48.95		173	45	28	103	5
246	75.9	16	57	1.78		276	170	33	135	27
90	1880.25	6.4	1140	26.7		145	30	55	30	7
920.4	1675.32	430.08	700.2	129.05		15	2	0	1	0
117	1024.65	0	525	12.46		248	72	48	140	6
109.8	48.99	21.12	24	1.78		272	10	14	8	5
7.8	3312	118.4	573	11.57		445	35	0	50	10
6.6	586.5	83.2	220.8	6.23		112	1485	22	490	37

8.4	834.9	92.16	234.6	7.565		103	8	29	24	10
10.8	53.82	154.88	93	1.78		117	15	30	30	5
42.6	293.94	68.48	175.2	3.56		861	48	14	290	43
18	258.75	64	138.6	6.942		278	12	27	15	8
37.8	182.85	69.12	88.8	5.34		38	5	0	2	1
106.2	46.23	44.8	79.8	0		953	86	0	80	83
218.4	20.7	5.12	66	14.24		734	12	5	180	38
59.4	315.33	81.92	216	16.91		439	41	21	163	15
219.6	177.33	35.2	58.8	13.35		254	13	99	135	6
102.6	34.5	77.184	57	6.586		363	17	66	120	7
77.16	114.54	116.48	124.8	3.56		1610	85	57	580	48
26.4	62.1	15.36	75	2.225		259	11	1484	200	6
180	11.04	7.68	34.2	2.67		233	58	383	10	2
72	72.45	62.72	97.2	4.45		32	3	0	2	0
-79.8	-28.8	-33.88	-68.64	-14.06		112	30	72	56	4
-513	-995.2	-263.34	-468			0	6	12	9	3
-61.2	-66.4	-23.1	-85.8	-14.06		0	7	2	6	2
-139.2	-20.8	-20.02	-202.8	-11.84		415	596	16	370	30
-120	-272	-79.31	-202.8	-33.3		264	13	10	19	10
-27	-22.4	-10.01	-80.34	-3.7		49	104	6	75	3
-102	-26.4	-40.04	-105.3	-19.98		435	10	28	133	10
-18	-44	-29.26	-23.4	-5.18		362	154	370	150	17
-1.2	0	-2.31	-0.78	0		353	35	36	86	9
-43.2	-38.4	-10.01	-109.2	-4.44		154	1375	210	660	18
-6	-11.2	-43.12	-6.24	-3.7		339	165	24	110	6
-21	0	-38.5	-39	-7.4		315	130	33	70	25
-891	-17.6	-53.9	-382.2	-27.38		390	195	75	210	5
-4.8	-23.2	-34.65	-18.72	-7.4		500	140	5	130	2
-9	-24	-26.18	-23.4	-3.7		290	50	5	30	1
-28.8	-11.2	-10.78	-226.2	-31.82		190	1347	5	6810	55
-7.2	-21.6	-32.34	-11.7	-5.92		410	110	25	95	2
-3	0	-4.62	-1.56	-0.74		150	2725	10	1900	30
-51.6	0	0	-62.4	-61.42		1534	2428	672	1167	145
-7.2	-4	-49.28	-140.4	-28.12		195	1485	0	875	14
-24.6	-16.8	-30.8	-127.14	-11.1		183	71	33	40	2
-7.8	-79.2	-10.01	-105.3	-4.44		13	4800	185	955	13
-10.2	-52.8	-30.8	-93.6	-5.18		11	850	130	368	7
-51	-45.6	-7.7	-452.4	-35.52		14	1210	144	391	8.5
-6.6	-1187.2	-388.08	-156	-4.44		18	78	242	155	2
-34.8	-306.4	-130.13	-7.8	-1.48		71	426	107	292	4
-1.8	0	-6.93	-1.56	0		30	375	100	231	7.8
-18	-57.6	-10.78	-43.68	-2.96		63	265	108	148	6
-3.6	-9.6	-11.55	-7.02	-2.22		177	67	70	133	0
-4.2	-1.6	-20.79	-4.68	-1.48		364	30	8	110	16
-357.6	-12.8	-31.57	-288.6	-22.2		99	457	128	360	19
-7.8	-8	-33.88	-14.82	-7.4		366	257	55	98	15
-62.4	-4.8	-5.39	-58.5	-2.22		171	50	120.6	95	7.4
-6	-22.4	-20.79	-103.74	-7.4		128.6	166	182	208	4

Interpretation of the results
Presenting the results and Visualising the results in graphics

The present study considereing 62 geothermal hot springs data focuses on spatially dependent multivariate geothermal data from two regions with diverse geologic-tectonic settings: the 2400 km-long arcuate belt in the tectonically active Extra-Peninsular Himalayan region and Late-Precambrian or Proterozoic mobile belts in the Central Highland in an otherwise stable landmass or shield of Peninsular India. It uses robust multivariate statistical methods, including Principal component analysis (PCA), exploratory factor analysis (EFA), correspondence factor analysis (CFA), and multiple regression analysis, to identify hidden patterns and determine the origin of geothermal hot springs (Amitabha Roy, 2024; 2023).





The model study distinguishes two statistically significant suites of fluid geochemistry: the overall acidic salt assemblage and concentration of Cl-HCO3-SO4-Na-F or chloride-rich water suggestive of a hydrothermal magmatic system operating in the geotherms of extra-peninsular India, and peninsular springs of alkaline K-Na-HCO3 bicarbonate-rich waters with low SO4-content and relatively higher contents of HCO3 compared to other anions SO4, Cl, and F suggestive of a non-magmatic origin. The graphical depiction via cluster analyses, the statistical measure of proximity and distance measurement from the origin of two-dimensional coordinate (factorial) axes, support the existence of two unique suites of geothermal geochemistry in two different geotectonic zones.

References

[1] Amitabha Roy, 2024. Geostatistics Applied To Fluid Geochemistry Of Geothermal Fields In Peninsular And Extra-Peninsular India. White Falcon Publishing, Chandigarh, India, 2024. Pp. 1-144. Isbn : 979-8-89222-356-0

[2] Amitabha Roy, 2023. A Comparative Statistical Study Of Geochemistry Of Geothermalffields Of Peninsular And Extra Peninsular India. J. Appl. Geol. & Geophys (Isor-Jagg), V. 11, Issue I, Ser. Ii, Pp. 32-44

[3] A.Roy, 1994. Gthermis – An Information Management And Analysis System For Geothermal Data Of India, A Field Season Report (1993-94).

[4] A.Roy,1981. Application Of Cluster Analysis In The Interpretation Of Geochemical Data From The Sargipalli Lead-Zinc Mine Area, Sundergarh District, Orissa (India). J.Geochemical Exploration. Elsevier Publ, Volume 14, Pp. 245-264.

[5] Jonathan Craig, 2013. Hot Springs And The Geothermal Energy Potential Of Jammu & Kashmir State, N.W. Himalaya, India.

[6] Mcqueen J, 1967. Some Methods For Classification And Analysis Of Multivariate Observations. Computer And Chemistry, 4, 257-272.Macqueen 1967)

[7] Ravi Shankar Et Al., 1991. Geothermal Atlas Of India, Gsi Spec Publ,