IOSR Journal Of Applied Geology And Geophysics (IOSR-JAGG)
E-ISSN: 2321–0990, P-ISSN: 2321–0982.Volume 12, Issue 2 Ser. I (Mar. – Apr. 2024), 01-08
Www.Iosrjournals.Org

Correspondence Factor Analysis (CFA) Of Multivariate Fluid Geochemistry Data From Indian Hot Springs

Amitabha Roy
Ex-Senior Director, Geological Survey Of India

Abstract
Correspondence analysis was conducted on a two-way contingency table with 62 rows and 9 columns (Primary dataset: A) representing fluid geochemical data from Indian hot springs. The analysis used the joint probability distribution of two random variables: individual observational samples (rows) and fluid geochemical variables (columns). This data set is really the total of two subsets: (B) Peninsula (25 rows and 9 columns) and (C) Extra-Peninsula (37 rows and 9 columns). The study generated factor loadings for individual samples and variables, which were shown as points on two-dimensional coordinate (factorial) axes with the same origin known as biplots, in order to find discrete geochemical domains defined by natural sample and variable groupings or clusters. The simultaneous changes in trace elements in these three data sets with sample locations appear to reflect broader trends in geothermal evolution in the region.
Keywords: *Correspondence factor analysis, Geochemistry, Biplots. Peninsula and Extra-Peninsula, Hot springs of India*

Date of Submission: 09-03-2024 Date of Acceptance: 19-03-2024

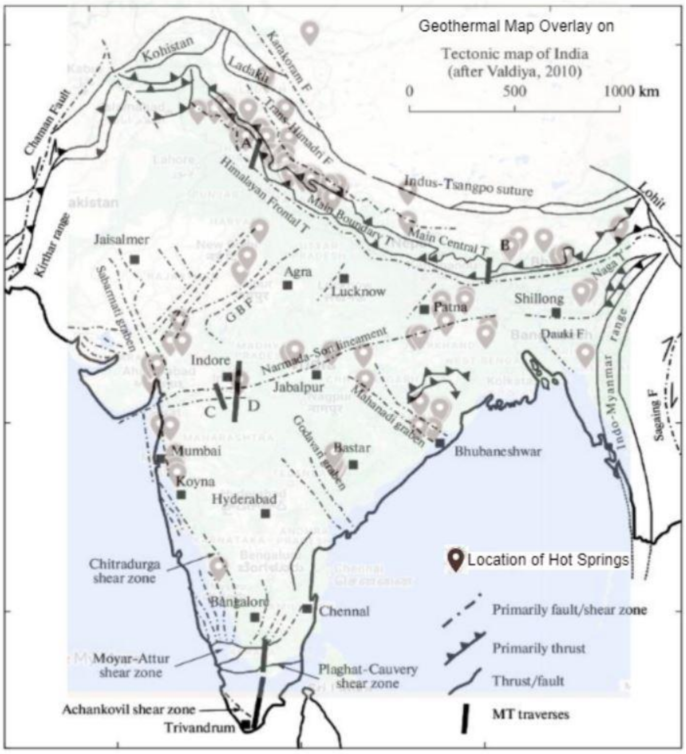
I. Introduction

The increasing demand for alternative renewable energy resources, such as geothermal, has led to a high interest in its exploration and exploitation. India has about 340 hot springs spread across the peninsular and extra-peninsular regions. The government of India constituted a 'Hot Spring Committee' in 1968 to examine the possibility of developing geothermal plants for power generation. The Central Electricity Authority (CEA) has associated itself with the UNDP geothermal project in India and the Puga and Parvati projects for the utilization of available geothermal resources for power generation (Jonathan Craig, 2013). The Geological Survey of India (GSI) has published a special publication titled "Geothermal Atlas of India" based on data compiled from all sources of information (Ravi Shankar et al. 1991). However, the lack of uniformity in data acquisition practices and manual handling of large amounts of data has made data storage, search, retrieval, and analysis laborious and cumbersome (A.Roy, 1994).

Study area and its geologic-tectonic settings

Two sets of multivariate geothermal data representing two spatially distinct regions of diverse geologic-tectonic settings, one from 2400 km long arcuate belt of tectonically active Extra-Peninsular Himalayan region and the other from Late-Precambrian or Proterozoic mobile belts in the Central Highland in otherwise stable landmass or shield of Penininsular India, were subjected to robust statistical techniques of Exploratory Factor Analysis followed by multiple regression analyses to find out the genesis of geothermal hot springs spread over these areas conspicuously associating with the respective tectonic zones of different degrees of severity (Amitabha Roy, 2023).

Fig, 1. Geothermal Map overlay on tectonic map of India



Both exploratory factor and multiple regression studies contribute to understanding the origins of these two fluid geochemistry suites. The model studies distinguish two statistically significant suites of fluid geochemistry: 1. the overall salt assemblage and concentration of Cl-HCO3-SO4-Na-F or chloride rich deep seated acidic waters suggestive of the existence of a hydrothermal magmatic system operating in the geotherms of Extra-Peninsular India; and 2. Peninsular springs of K-Na-HCO3 bicarbonate rich alkaline waters with low SO4-content and relatively higher contents of HCO3 compared to other anions SO_4, Cl, and F suggestive of a non-magmatic origin.

In the present study, correspondence factor analysis, a multivariate statistical technique of proximity and distance measure from the origin of two-dimensional coordinate (factorial) axes, was performed using XLSTAT software to identify associations or oppositions between observation samples (rows) and multivariate fluid geochemical data (columns), calculating their contribution to total inertia for each factor. The projection of the rows and columns onto the same set of factorial axes with the same origin enables the development of two-dimensional graphs, which aid in the interpretation of the results. To be more specific, the resultant graph is an overlay of row coordinates over column coordinates, or vice versa.

Computational strategy

Correspondence analysis is a statistically based geometric technique that displays the rows and columns of a two-way contingency table as points in a two-dimensional vector space (Benzekri, 1973; David et al., 1977; Davis, 1986; Teil, 1975). In this analysis, the contingency table is looked upon as a joint probability distribution of two random variables, namely, individual observations or samples (i = 1, 2, 3,..., N) and variables (j = 1, 2, 3,..., M). The raw data matrix (X_{ij}) is converted into a matrix of joint probability (P_{ij}) of occurrence by dividing each cell entry by the sum of the data values in all rows and columns, i.e.

$$P_{ij} = X_{ij}/K \text{ where } K = \sum_{i=1}^{N} \sum_{j=1}^{M} X_{ij}$$

Sum of the probabilities P_{ij} in all rows and columns is given by

$$\sum_{i=1}^{N} \sum_{j=1}^{M} P_{ij} = 1$$

the row-totals of each row

$$P_i = \sum_{j=1}^{M} P_{ij} = K_i/K$$

and the column-totals of each columns

$$P_j = \sum_{i=1}^{N} P_{ij} = K_j/K$$

give the marginal probabilities of each sample and variable respectively

The joint probability distribution matrix generated from the contingency table (Xij) is then turned into a square, symmetric matrix for the computation of eigenvalues and eigenvectors, from which factor loadings for samples and variables are extracted in the usual way (Davis, 1986). The factor loadings are then utilized to represent samples (I) and variables (J) simultaneously on factorial axes. The correspondence of the i-th sample and the j-th variable on the q-th factorial axis is given by

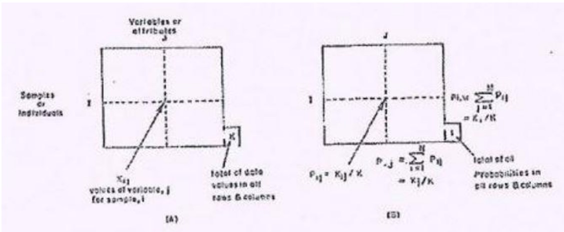


Fig 2. Geometric representation of transformation of (A) two-dimensional contingency table (X_{ij}) having M(j) variables and N observations or samples (i) into (B) joint probability distribution matrix (P_{ij}) by correspondence analysis

$$Sq(i) = \frac{1}{\sqrt{\lambda q}} \sum_{j=1}^{N} Vq(j) \frac{P_{ij}}{p_{i.}}$$

$$Vq(j) = 1/\sqrt{\lambda q} \sum_{i=1}^{n} Sq(i) \frac{P_{ij}}{P_j}$$

Where λq = eigenvalues or inertia of factorial axis q
$S_{q(i)}$= abscissae or loadings of i on factorial axis q
$V_{q(i)}$ =abscissae or loadings of j on factorial axis q
$\frac{P_{ij}}{P_i}$ = conditional probability of drawing a variable given that it belongs to sample i
$\frac{Pij}{P.j}$ =condititional probability of drawing a sample given that it belongs to variable of type j

Correspondence analysis dataset (A): a two-way contingency table with 62 rows and 9 columns

The fundamental input data, or in this example, the two-way contingency table, consists of 62 rows or records and 9 columns or categories. Of the 62 rows, 37 represent fluid geochemical data from hot springs in the Extra-Peninsula (Himalayan), whereas 25 reflect data from Indian Peninsula locations.
Distance: Chi-square
Significance level (%): 5
Filter factors Maximum number: 5
Rotation: Varimax (Kaiser normalization) / Based on columns / Number of factors = 2

Test of independence between the rows and the columns:

Chi-square (Observed value)	67098.9188
Chi-square (Critical value)	540.499
DF	488
p-value	**<0.0001**
alpha	0.05

Test interpretation:
H0: The rows and the columns of the table are independent.
Ha: There is a link between the rows and the columns of the table.

As the computed p-value is lower than the significance level alpha=0.05, one should reject the null hypothesis H0, and accept the alternative hypothesis Ha.

Total inertia: 0.86

Eigenvalues and percentages of inertia:

	F1	F2	F3	F4	F5	F6	F7	F8
Eigenvalue	0.375	0.240	0.160	0.035	0.025	0.013	0.007	0.004
Inertia %	43.624	27.903	18.601	4.061	2.914	1.550	0.837	0.509
Cumulative %	43.624	71.528	90.129	94.190	97.104	98.654	99.491	100.000

Fig. 3. Scree plot showing the percentages of inertia Captured by the new dimensions generated by CA

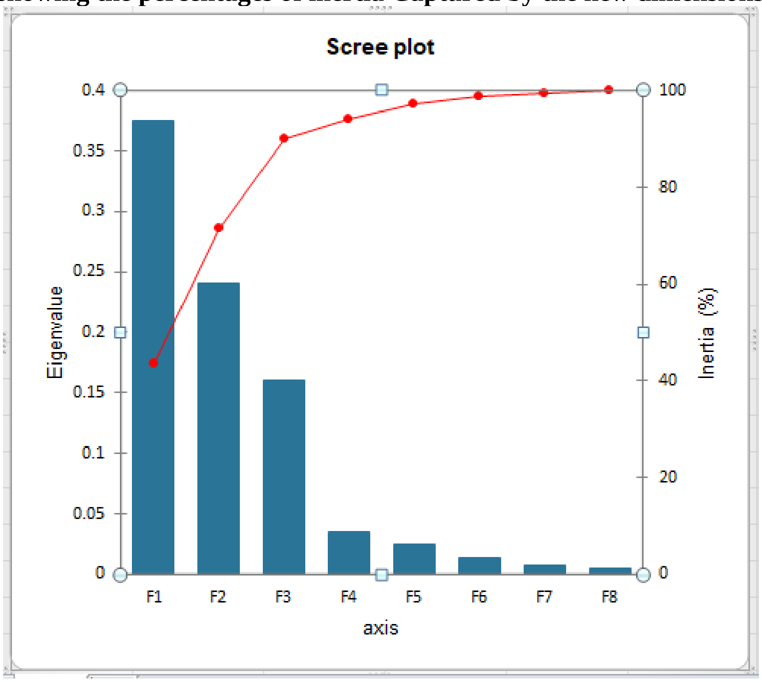


Table 1. A two-way contingency table

	HCO3 mg/L	Cl mg/L	SO4 mg/L	Ca mg/L	Mg mg/L	Na mg/L	K mg/L	F mg/L	B mg/L
1	300.000	163.000	62.000	14.000	5.000	210.000	13.000	12.000	5.000
2	170.000	133.000	36.000	44.000	15.000	88.000	19.000	0.800	33.000
3	490.000	855.000	1244.000	342.000	87.000	600.000	109.000	3.600	138.000
4	210.000	102.000	83.000	30.000	15.000	110.000	19.000	1.200	25.000
5	342.000	232.000	26.000	26.000	1.000	260.000	16.000	10.000	10.000
6	303.000	200.000	340.000	103.000	11.000	260.000	45.000	6.000	13.000
7	173.000	45.000	28.000	13.000	2.000	103.000	5.000	10.000	3.000
8	276.000	170.000	33.000	52.000	12.000	135.000	27.000	3.000	10.000
9	145.000	30.000	55.000	38.000	13.000	30.000	7.000	1.000	3.000
10	15.000	2.000	0.000	3.000	1.000	1.000	0.000	0.200	0.000
11	248.000	72.000	48.000	13.000	2.000	140.000	6.000	5.000	3.000

12	272.000	10.000	14.000	56.000	24.000	8.000	5.000	0.400	0.000
13	445.000	35.000	0.000	50.000	52.000	50.000	10.000	1.200	0.000
14	112.000	1485.000	22.000	70.000	13.000	490.000	37.000	1.600	19.000
15	103.000	8.000	29.000	45.000	44.000	24.000	10.000	0.700	2.000
16	117.000	15.000	30.000	34.000	3.000	30.000	5.000	1.600	0.000
17	861.000	48.000	14.000	14.000	99.000	290.000	43.000	3.000	5.000
18	278.000	12.000	27.000	42.000	26.000	15.000	8.000	0.500	1.000
19	38.000	5.000	0.000	6.000	7.000	2.000	1.000	0.400	0.000
20	953.000	86.000	0.000	0.000	47.000	80.000	83.000	0.000	0.000
21	734.000	12.000	5.000	64.000	10.000	180.000	38.000	2.000	1.000
22	439.000	41.000	21.000	40.000	23.000	163.000	15.000	4.000	2.000
23	254.000	13.000	99.000	13.000	0.000	135.000	6.000	12.500	2.800
24	363.000	17.000	66.000	40.000	8.000	120.000	7.000	10.000	2.000
25	1610.000	85.000	57.000	10.000	2.000	580.000	48.000	10.000	8.000
26	259.000	11.000	1484.000	504.000	22.000	200.000	6.000	2.500	1.000
27	233.000	58.000	383.000	169.000	28.000	10.000	2.000	0.200	0.000
28	32.000	3.000	0.000	9.000	0.000	2.000	0.000	0.400	0.000
29	112.000	30.000	72.000	14.000	1.000	56.000	4.000	6.000	1.000
30	0.000	6.000	12.000	15.000	3.000	9.000	3.000	1.000	0.900
31	0.000	7.000	2.000	27.000	5.000	6.000	2.000	0.200	0.900
32	415.000	596.000	16.000	41.000	21.000	370.000	30.000	3.000	8.000
33	264.000	13.000	10.000	44.000	18.000	19.000	10.000	0.300	0.000
34	49.000	104.000	6.000	7.000	1.000	75.000	3.000	7.000	1.000
35	435.000	10.000	28.000	27.000	11.000	133.000	10.000	2.100	0.000
36	362.000	154.000	370.000	127.000	19.000	150.000	17.000	1.000	0.000
37	353.000	35.000	36.000	54.000	5.000	86.000	9.000	0.000	0.000
38	154.000	1375.000	210.000	204.000	88.000	660.000	18.000	0.700	0.000
39	339.000	165.000	24.000	82.000	16.000	110.000	6.000	0.400	0.900
40	315.000	130.000	33.000	110.000	12.000	70.000	25.000	0.000	0.000
41	390.000	195.000	75.000	65.000	40.000	210.000	5.000	0.000	0.100
42	500.000	140.000	5.000	70.000	40.000	130.000	2.000	1.000	1.200
43	290.000	50.000	5.000	60.000	20.000	30.000	1.000	0.300	1.200
44	190.000	1347.000	5.000	390.000	250.000	6810.000	55.000	0.000	0.000
45	410.000	110.000	25.000	45.000	15.000	95.000	2.000	0.000	0.000
46	150.000	2725.000	10.000	105.000	40.000	1900.000	30.000	0.200	3.000
47	1534.000	2428.000	672.000	9.000	8.000	1167.000	145.000	0.000	0.000
48	195.000	1485.000	0.000	90.000	40.000	875.000	14.000	0.000	0.000
49	183.000	71.000	33.000	40.000	21.000	40.000	2.000	0.000	0.000
50	13.000	4800.000	185.000	186.000	10.000	955.000	13.000	0.000	0.400
51	11.000	850.000	130.000	170.000	0.100	368.000	7.000	2.000	0.400
52	14.000	1210.000	144.000	348.000	0.200	391.000	8.500	7.200	0.000
53	18.000	78.000	242.000	40.000	15.000	155.000	2.000	2.500	0.000
54	71.000	426.000	107.000	32.000	6.000	292.000	4.000	1.500	1.000
55	30.000	375.000	100.000	56.000	1.800	231.000	7.800	4.000	0.400
56	63.000	265.000	108.000	80.000	44.000	148.000	6.000	0.100	0.000
57	177.000	67.000	70.000	3.000	1.000	133.000	0.000	3.000	0.500
58	364.000	30.000	8.000	35.000	3.000	110.000	16.000	0.300	0.000
59	99.000	457.000	128.000	42.000	2.000	360.000	19.000	0.500	0.000
60	366.000	257.000	55.000	96.000	70.000	98.000	15.000	0.200	0.000
61	171.000	50.000	120.600	50.000	7.900	95.000	7.400	4.000	0.000
62	128.600	166.000	182.000	20.000	13.400	208.000	4.000	5.000	0.000

Results after the Varimax rotation (Kaiser normalization):

Rotation matrix:

	D1	D2
D1	-0.931	-0.364
D2	-0.364	0.931

Percentage of variance after Varimax rotation:

	D1	D2	F3	F4	F5
Variability	41.541	29.987	18.601	4.061	2.914
Cumulative	41.541	71.528	90.129	94.190	97.104

Table 2.

Standard coordinates (rows)		
after varimax rotation		
	D1	D2
1	0.634	-0.214
2	0.448	0.186
3	-0.296	1.680
4	0.601	0.445
5	0.554	-0.492
6	0.153	1.178
7	0.990	-0.198
8	0.669	-0.162
9	1.010	0.779
10	1.820	-0.226
11	0.950	-0.146
12	1.989	0.037
13	1.977	-0.412
14	-0.978	-0.524
15	1.152	0.587
16	1.178	0.521
17	1.775	-0.631
18	1.915	0.101
19	1.851	-0.290
20	2.197	-0.617
21	1.969	-0.514
22	1.541	-0.415
23	1.096	0.449
24	1.471	0.103
25	1.771	-0.616
26	-0.311	3.596
27	0.259	2.624
28	1.829	-0.068
29	0.671	0.925
30	-0.371	1.762
31	-0.260	1.262
32	0.092	-0.585
33	1.976	-0.122
34	-0.183	-0.526
35	1.791	-0.374
36	0.349	1.522

Contribution(rows)		
after Varimax rotation:		
	D1	D2
1	0.004	0.000
2	0.001	0.000
3	0.004	0.140
4	0.003	0.002
5	0.004	0.003
6	0.000	0.023
7	0.005	0.000
8	0.004	0.000
9	0.004	0.003
10	0.001	0.000
11	0.006	0.000
12	0.020	0.000
13	0.032	0.001
14	0.028	0.008
15	0.005	0.001
16	0.004	0.001
17	0.056	0.007
18	0.019	0.000
19	0.003	0.000
20	0.077	0.006
21	0.052	0.004
22	0.023	0.002
23	0.008	0.001
24	0.018	0.000
25	0.097	0.012
26	0.003	0.413
27	0.001	0.078
28	0.002	0.000
29	0.002	0.003
30	0.000	0.002
31	0.000	0.001
32	0.000	0.007
33	0.019	0.000
34	0.000	0.001
35	0.027	0.001
36	0.002	0.036

37	1.548	-0.058
38	-0.781	-0.040
39	0.879	-0.171
40	0.953	0.105
41	0.709	-0.081
42	1.342	-0.429
43	1.639	-0.182
44	-0.482	-1.032
45	1.356	-0.293
46	-0.937	-0.771
47	-0.053	-0.016
48	-0.785	-0.704
49	0.992	0.162
50	-1.316	-0.383
51	-1.050	0.106
52	-1.050	0.205
53	-0.617	2.109
54	-0.728	0.002
55	-0.857	0.192
56	-0.496	0.546
57	0.632	0.152
58	1.697	-0.477
59	-0.624	0.010
60	0.670	0.009
61	0.548	1.028
62	-0.167	0.825

37	0.018	0.000
38	0.021	0.000
39	0.007	0.000
40	0.008	0.000
41	0.006	0.000
42	0.021	0.002
43	0.016	0.000
44	0.027	0.123
45	0.017	0.001
46	0.056	0.038
47	0.000	0.000
48	0.021	0.017
49	0.005	0.000
50	0.137	0.012
51	0.022	0.000
52	0.030	0.001
53	0.003	0.032
54	0.006	0.000
55	0.008	0.000
56	0.002	0.003
57	0.002	0.000
58	0.021	0.002
59	0.006	0.000
60	0.006	0.000
61	0.002	0.007
62	0.000	0.006

II. Interpreting The Results

The findings of correspondence analysis are visually interpreted using scree plots and biplots, as well as statistically using output statistics.

Scree plot

A scree plot, illustrated in Figure 3, may be used to compare the percentages of total inertia explained by the new CA dimensions. In our case, the first dimension accounts for 89.4% of the inertia, whereas the second accounts for 10.19%. The first two dimensions collectively account for 99.5% of total inertia.

Biplot

The row and column variables are then projected into the first two dimensions and graphically explored using a biplot. The biplot depicts the first two dimensions on the x and y axes, respectively. Transition formulae between the coordinates of the row, column, and additional variables can be used to display them along the same axes. Proximity in the feature space suggests a favorable connection.

III. Discussion

Graphical representation of a contingency table

A few crucial factors to remember while analyzing correspondence analysis include:

1) Use raw data to verify findings,
2) The farther things are from their origin, the more discriminating they are,
3) The closer anything is to its origin, the less distinct it is,
4) The more variance explained, the less insights will be lost,
5) Proximity between row labels suggests similarity;
6) Proximity between column labels shows similarity; and
7) If a tiny angle connects a row and column label to the origin, they are most likely related,
8) If a row and column label form a 90-degree angle with the origin, they are most likely unrelated,
9) If a row and column label are on opposing sides of the origin, they are most likely negatively connected, and
10) The further a point from the origin, the greater its positive or negative relationship.

The primary dataset, the two-way contingency table (Table-1), contains 62 rows and 9 columns. It is divided into two subsets: 25 rows and 9 columns for Peninsular India and 37 rows and 9 columns for Extra-Peninsular India. While the findings of the correspondence analysis of the main dataset has been replicated in their entirety, only the results of biplots for the two subsets have been provided with the goal of matching the overall results.

Primary dataset: contingency table with 62 rows and 9 columns:
Looking at the biplots, three clusters are discernible: a) predominantly alkaline HCO3- Mg-K-F further right from the origin in close proximity and aligning with the factorial D1 axis; b) acidic to neutral group SO4-B-Ca further north of the origin aligning with the vertical D2 factorial axis and forming a 90-degree angle in respect of group (a) with the origin, thus uncorrelated; and c) acidic to neutral group Cl-Na making a tiny angle that connects a row and column label to the the left of the origin close to the factorial D1 axis.

Subset Peninsula with 25 rows and 9 columns:
Biplots show two different opposing groups north and south of the origin:
a) the generally alkaline group HCO3-Mg-Ca-Na-K-B, which is negatively related with
b) the mostly acidic group Cl-SO4-F. south of the origin.

Subset Extra-Peninsula with 37 rows and 9 columns:
Biplots show two opposing groups north and south of the origin:
a) the typically acidic
Cl-SO4-B-Na north of the origin, which is negatively connected to
b) the largely alkaline
HCO3-Mg-Ca-K-F south of the origin.

When the biplot results of correspondence analysis are compared to the three datasets stated above, it is evident that the more data there is, the more probable it is that any good summary may neglect vital information, as is the case with the major dataset (A). Surprisingly, biplots of subsets (B) and (C) swap places across the origin of two-dimensional coordinate (factorial) axes.

References

[1]. Amitabha Roy, 2023. Geostatistics As Applied To The Fluid Geochemistry Of Indian Hot Springs. Jour. Applied Geology And Geophysics, V. 11, Issue 4, Ser. Ii, Pp 1-37
[2]. A.Roy, 1994. Gthermis – An Information Management And Analysis System For Geothermal Data Of India, A Field Season Report (1993-94).
[3]. A.Roy, A.K. Saha And S.N. Sarkar,1994. Factorial Correspondence Analysis Of Spatially Related Multi-Metallic Data Along Gangpur-Singhbhum Metallotectonic Belt. Jour. Geological Society Of India, V. 43 (Apr.). Pp. 395-406
[4]. Benzekri, J.P, 1973. I'analyss Des Donnaes, Tome Ii, Paris, Pp. 619
[5]. David, M. Et Al., 1977. Correspondence Analysis. Quart. Colorado Sch Mines, V.72, Pp. 11-57
[6]. Davis, J.C.,1986. Statistical Data Analysis In Geology. John Wiley, N.Y., Pp. 646
[7]. Jonathan Craig, 2013. Hot Springs And The Geothermal Energy Potential Of Jammu & Kashmir State, N.W. Himalaya, India.
[8]. Kaiser, H.F. (1958). The Varimax Criterion For Analytic Rotation In Factor Analysis. Psychometrika, 23, Pp.187-200.
[9]. Ravi Shankar Et Al. Geothermal Atlas Of India. Gsi Spec Publ, (1991)
[10]. Teil, H, 1975. Correspondence Factor Analysis:An Outlining Of Method. Mathematical Geology, 7, Pp. 3-12