IOSR Journal Of Applied Geology And Geophysics (IOSR-JAGG)
E-ISSN: 2321–0990, P-ISSN: 2321–0982.Volume 13, Issue 2 Ser. I (Mar. – Apr. 2025), 21-26
www.Iosrjournals.Org

Latent Class Clustering And Profile Analysis To Uncover Hidden Patterns Within The Mixed Observed Distribution Of Multivariate Fluid Geothermal Geochemistry Data In Indian Hot Springs

Amitabha Roy
Ex-Senior Director, Geological Survey Of India

Abstract
This research paper explores the application of Latent Profile Analysis (LPA) focusing on identifying hidden patterns in the mixed observed distribution of multivariate fluid geothermal geochemistry in Indian hot springs using latent class clustering. The study uses robust multivariate statistical techniques to spatially dependent multivariate geothermal geochemistry data from hot springs located in two regions with different geologic-tectonic settings: a 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. The study addresses the complexities of selecting significant latent classes from a multitude of factors, utilising graphical representations to illustrate the results of profile analysis, which identified five clusters of geochemical characteristics.By employing Latent Class Analysis (LCA), distinct subgroupings emerged, revealing three significant profiles: HCO3-Cl-SO4, Ca-Mg-Na-K, and F-B, each associated with varying geochemical properties. The findings highlight that the red group (F-B) exhibits weakly acidic properties, while the blue group (Ca-Mg-Na-K) is alkaline, and the green group (HCO3-Cl-SO4) is acidic. The comparison of LCA with other statistical methods, such as traditional cluster analysis and factor analysis, underscores LCA's effectiveness in analyzing categorical data and extracting meaningful patterns in heterogeneous geological settings. Ultimately, this study contributes to a deeper understanding of the geochemical diversity of Himalayan, and other Penisular hot springs, emphasizing the utility of LCA in environmental and geological research.

Keywords: *Latent Class Cluster (LCC), Profile plots, k-mean cluster analysis, factor analysis, geothermal geochemistry, hot springs, Peninsula, extra- Peninsula, India*

Date of Submission: 24-03-2025 Date of Acceptance: 04-04-2025

I. Background Knowledge

There are around 340 hot springs located throughout India, including the Peninsular and Extra-Peninsular regions. Schlagintweit compiled the first list of hot springs in India in 1852. In 1991, the Geological Survey of India produced a 'Geothermal Atlas of India', and the Indian government established a 'Hot Spring Committee' to investigate the development of geothermal power plants. Projects in the Puga Valley and Parvati aim to harvest 5000 MWh of geothermal energy, which is enough to run a 20 MWe power plant (Jonathan Craig, 2013). GTHERMIS is a computerised geothermal database system developed by the GSI (A. Roy, 1994).

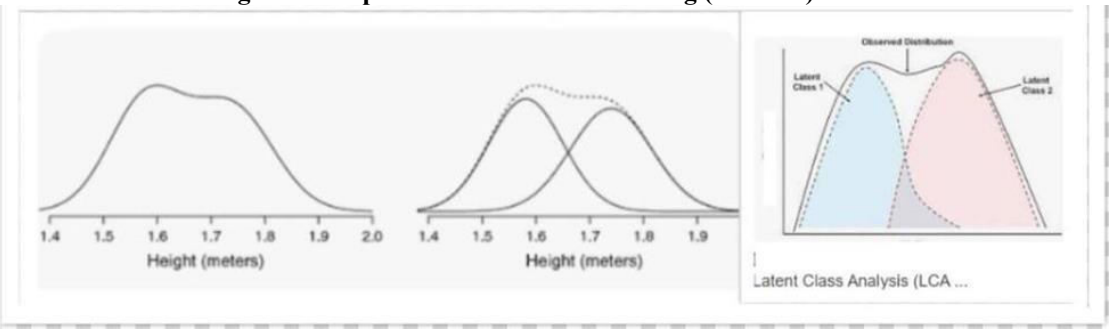
An extensive study using a range of robust multivariate statistical techniques was carried out to spatially dependent multivariate geothermal geochemistry data from hot springs located in two regions with different geologic-tectonic settings: a 2400 km-long arcuate belt in the tectonically active Extra-Peninsular Himalayan region, and late-Precambrian or Proterozoic mobile belts in the Central Highland in an otherwise stable landmass or shield of Peninsular India. The goal was to uncover hidden patterns in the geothermal geochemistry of hot springs. (Amitabha Roy, 2024).

II. Latent Class Clustering Model Analysis

Latent class modelling is an effective strategy for generating meaningful segments that differ from response patterns associated with categorical (correspondence analysis) or continuous variables (factor analysis or k-means cluster analysis). Latent class cluster analysis (LCA) is a statistical technique for grouping multivariate discrete data. It is assumed that the data are derived from a mixture of discrete distributions of continuous, mutually exclusive, and exhaustive independent or uncorrelated variables (i.e., indicators) on a set

of (categorical) measured variables. LCA also assumes that there are unobserved heterogeneity and existence of subgroups or latent profiles that form patterns of responses to indicator variables.

Fig. 1. Conceptual Latent Class Clustering (Mixture) Model



Computational strategy

f (yi|θ) =K∑k=1πkfk(yi|θk), where

yi denotes an object's scores, mixed continuous (as against categorical) indicators variables, dependent variables, or items on a set of observed variables, K is the number of clusters,and πk denotes the prior probability of belonging to latent class or cluster k or, equivalently, the size of cluster k. As can be seen, the distribution of yi given the model parameters θ, f (yi|θ), is assumed to be a mixture of classes-geochemical variables,fk(yi|θk). These continuous variables are assumed to be normally distributed within latent classes. The fact that each class has its own separate sret of means,variances, and covariances for each latent class means that the y variables may be correlated with clusters, as well as that these correlations may be cluster specific.

III. The dataset used in the study and its earlier visualization

Fig. 2. Google Map Showing Locations Of Hot Springs (Right) And Geothermal Geochemistry Dataset With Rows (Observations) =82 And Columns (Variables) =9 (Left)

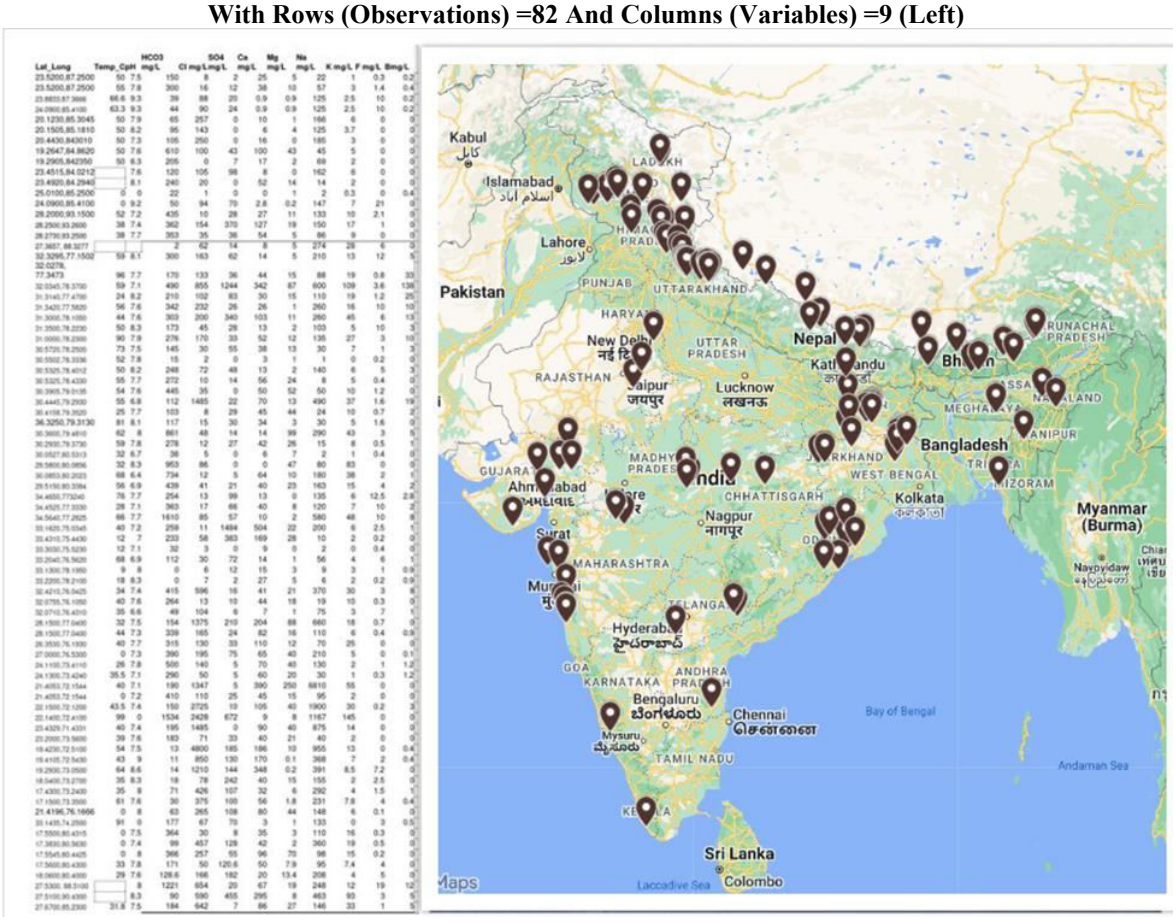


Fig. 3. Varimax Factor Analysis

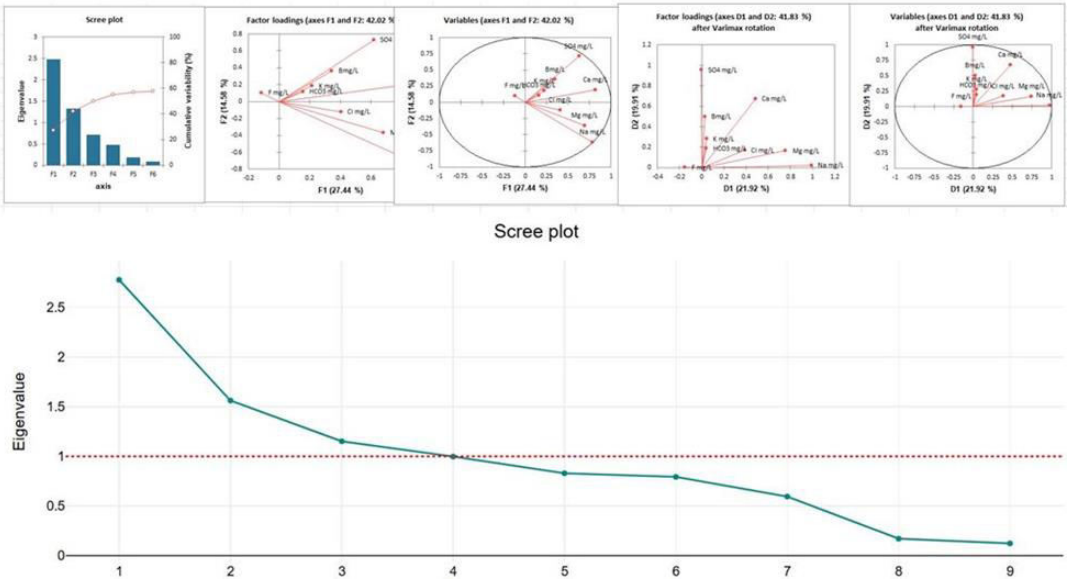


Fig.4. K-Mean Cluster Analysis

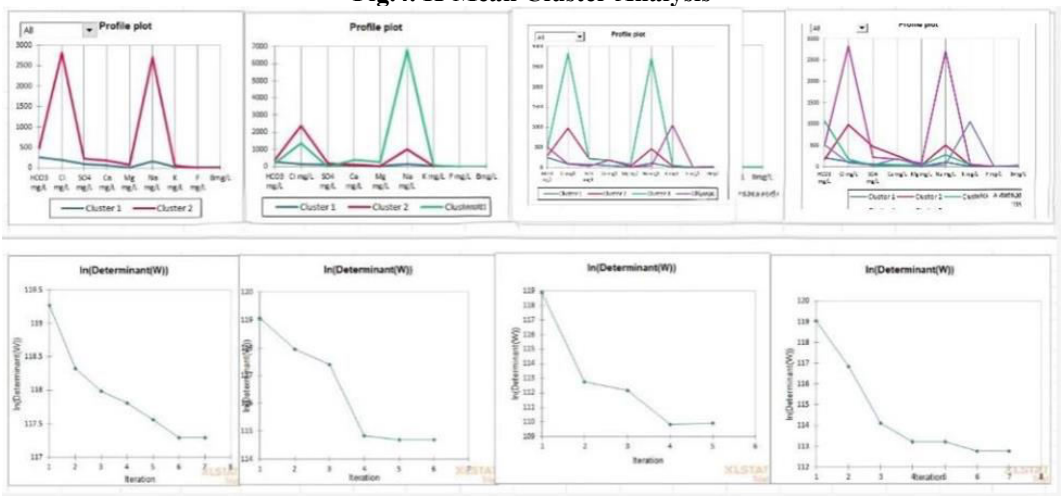


Fig. 5. 3D Model Of K-Mean Cluster Analysis

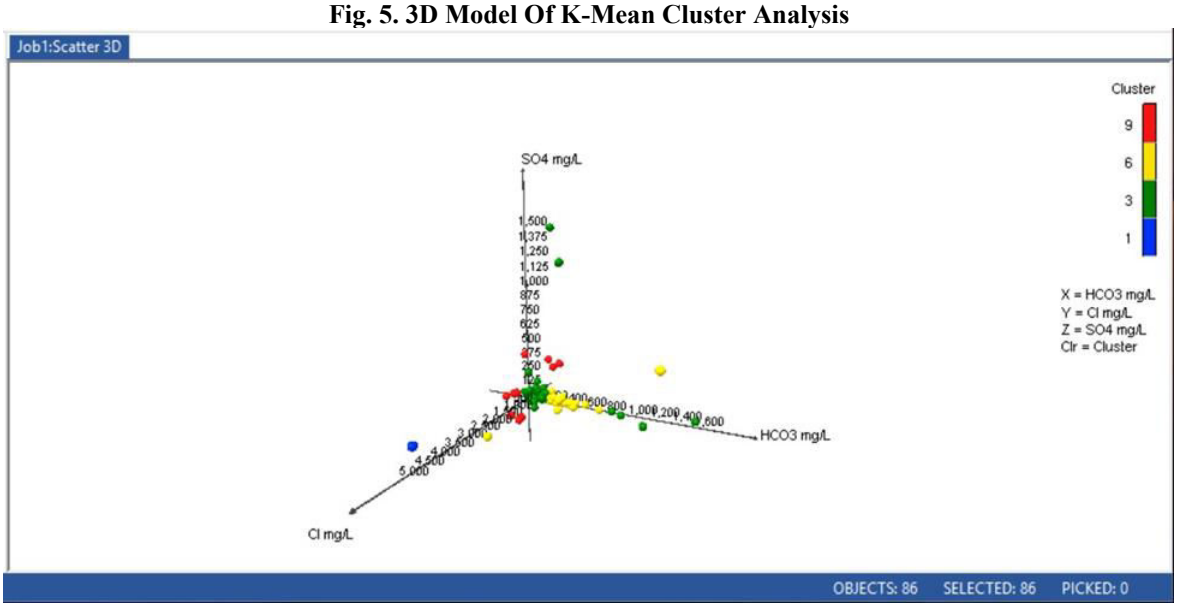
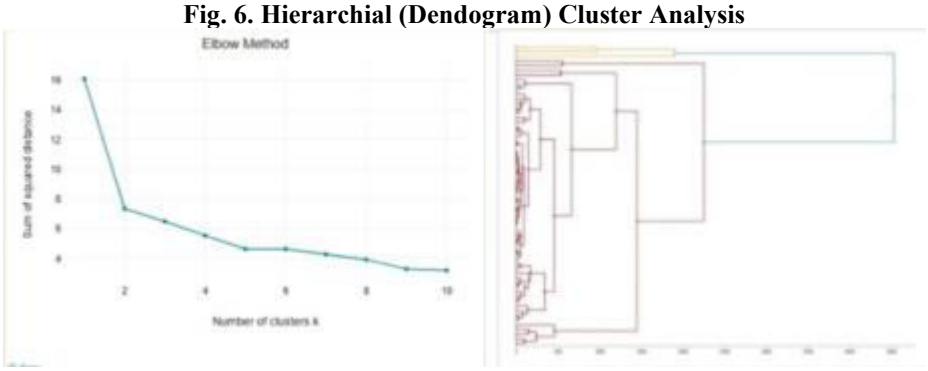


Fig. 6. Hierarchial (Dendogram) Cluster Analysis



IV. Interpretation Of Results

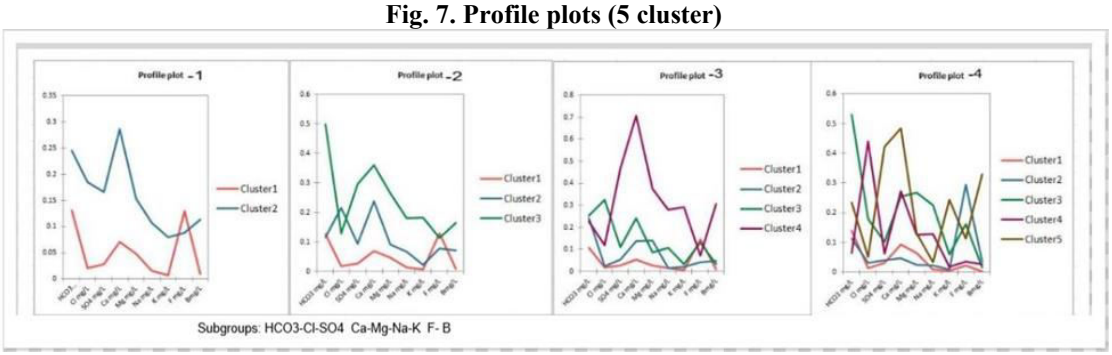
Presenting the results and visualizing the results in Latent Profile Plots

Table. 1

Summary statistics (5 Clusters):						
Number of cases	82					
Number of parameters	94					
Seed (random numbers)	123456789					
Best seed	613427					
EM iterations	2					
Log-posterior	-3696					
L^2	7157					
Final convergence value	0					
Newton-Raphson iterations	4					
Log-posterior	-3696					
L^2	7157					
Final convergence value	0					
Log-likelihood statistics (5 Clusters):						
Log-likelihood(LL)	-3578					
Log-prior	-118					
Log-posterior	-3696					
BIC(LL)	7571					
AIC(LL)	7345					
AIC3(LL)	7439					
CAIC(LL)	7665					
SABIC(LL)	7274					
Classification statistics (5 Clusters):						
Classification errors	0					
Reduction of errors (Lambda)	1					
Entropy R^2	1					
Standard R^2	1					
Classification Log-likelihood	-3582					
Entropy	4					
CLC	7164					
AWE	8274					
ICL BIC	7578					
Classification table - Modal (5 Clusters):						
Probabilistic\Modal	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Total
Cluster1	41	0	0	0	0	41
Cluster2	1	23	0	0	0	24
Cluster3	0	0	6	0	0	6
Cluster4	0	0	0	6	0	6
Cluster5	0	0	0	0	5	5
Total	42	23	6	6	5	82
Classification table – Proportional (5 Clusters)						
Probabilistic\Modal	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Total
Cluster1	40	1	0	0	0	41
Cluster2	1	23	0	0	0	24

Cluster3	0	0	6	0	0	6
Cluster4	0	0	0	6	0	6
Cluster5	0	0	0	0	5	5
Total	41	24	6	6	5	82
Profile (5 Clusters):						
Variables	Categories	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Cluster size		0	0	0	0	0
HCO3 mg/L	HCO3 mg/L	226	182	854	106	379
Cl mg/L	Cl mg/L	68	157	861	2113	243
SO4 mg/L	SO4 mg/L	46	61	150	93	628
Ca mg/L	Ca mg/L	47	24	128	137	244
Mg mg/L	Mg mg/L	16	6	67	32	33
Na mg/L	Na mg/L	69	164	1539	872	228
K mg/L	K mg/L	6	10	63	20	256
F mg/L	F mg/L	0	6	3	1	2
Bmg/L	Bmg/L	0	5	2	4	46

Fig. 7. Profile plots (5 cluster)



V. Visualizing Latent Profile Analysis (LPA)

The most baffling issue of all of these methodologies, choosing a few significant factors/clusters/latent class clusters from many insignificant ones is a tricky task. The usual criterion is to use the inflection point of the scree or elbow curve. In the absence of a clear inflection point as in factor analysis, the eigenvalue > 1 or the factor loadings for each variable ≥ 0.6 is used (Awang, 2014).

The results of the Table.1. Profile (5 clusters) table are shown graphically. The X axis displays the item number or variable name, while the Y axis represents probability. The three classes are denoted by three different coloured lines: red, blue, and green.

Now, utilising LCC, it would safely choose a three-profile solution consisting of subgroups: 1) HCO3-Cl-SO4, 2) Ca-Mg-Na-K, and 3) F-B. Profile plots highlight some intriguing variances. Clearly, the red group F-B (Fluorine-Boron) is weak or Lewis acid, the blue group Ca-Mg-Na-K is alkaline, and the green group (HCO3-Cl-SO4) is acidic [(Fig. 7(2)].

When comparing the effectiveness of LCA and factor analysis, it has been discovered that the latent class cluster corresponds more or less to the latent unrotated PCA factors than to the rotated varimax factor. This is consistent with the question "Why is factor rotation (overmaximization) usually advised, even when it obscures general factors?" (A. Fog, 2014).

Latent profile analysis also provides an intriguing conclusion. Tectonically active extra-peninsular (Himalayan) springs are magmatic deep-seated hydrothermal manifestations, a phenomenon in which magma gradually degasses in decreasing order of solubility, i.e., $CO_2 > SO_2 > HCl > HF$, i.e., CO_2 - first until HF -last (Giggenbach, 1987).

VI. LCA Vs. Cluster Analysis And Factor Analysis

Latent Class Analysis (LCA) is a method used to group data into clusters, similar to cluster analysis. It focuses on the structure of groups rather than correlations between variables, which is the emphasis of Factor Analysis. One key difference is that LCA deals with discrete latent categorical variables that follow a multinomial distribution, while Factor Analysis works with continuous latent variables that have a normal distribution. LCA emerged from social sciences, where many variables are not continuous.

Latent class segmentation and k-means clustering are techniques for grouping data points, with different methodologies and advantages. Latent class segmentation is effective for categorical data and provides

clear insights into data structure. It allows for flexible distribution assumptions and is a model-based approach. LCMs can incorporate missing data more naturally through maximum likelihood estimation techniques.

VII. Conclusion

In conclusion, the evaluation of various statistical data reduction strategies in the context of multivariate fluid geothermal geochemistry in Indian hot springs highlights the efficacy of latent class clustering (LCA) in identifying hidden patterns within disparate geological environments. The study encompassed 82 hot spring samples, utilizing robust multivariate techniques to capture spatially dependent geochemical data from regions with varying tectonic settings. Through LCA, distinct geochemical profiles were revealed, categorizing the springs into three significant out of five insignificant clusters based on indicator variables such as HCO3, Cl, and SO4 concentrations. Notably, the results suggest that the Himalayan hot springs display characteristics indicative of magmatic hydrothermal activity, underscoring the variability in geochemical compositions linked to their geological contexts. Furthermore, the comparison between LCA and traditional cluster analysis or factor analysis illustrates the unique advantages of LCA in handling categorical data and uncovering meaningful latent groupings, which could be pivotal for future geothermal resource management strategies. Overall, this comprehensive analysis not only advances our understanding of geothermal geochemistry in Indian hot springs but also serves methodological template for similar studies in geoscience.

References

[1] A.Roy, 1994. GTHERMIS – An Information Management And Analysis System For Geothermal Data Of India, A Field Season Report (1993-94)
[2] Amitabha Roy, 2024. Geostatistics Applied To Fluid Geochemistry Of Geothermal Fields In Peninsular And Extra-Peninsular India. White Falcon Publishing, Chandigarh, India, 2024. Pp. 1-142. ISBN: 979-8-89222-356-0
[3] Amitabha Roy, 2023. Geostatistics As Applied To Fluid Geochemistry Of Indian Hot Springs. J. Appl. Geol. & Geophys (ISOR-JAGG), V.11, Issue 4, Ser. II, Pp. 01-37
[4] Amitabha Roy, 2024. Data Visualization To Understand How Data Is Structured Using K-Means And Hierarchial Cluster Analysis. J. Appl. Geol. & Geophys (ISOR-JAGG), V.12, Issue 6, Ser. I, Pp. 01-05
[5] A.Fog, 2014. Why Is Factor Rotation Always Recommended, Though It Obscures General Factors?, J. Ross-Cultural Research
[6] Awang, Z, 2014. Research Methodology And Data Analysis ...
[7] Belany Bray, 2021. Introduction To Latent Classes And Latent Profile Analysis. Jour. Social Sciences
[8] Jonathan Craig, 2013. Jonathan Craig, 2013. Hot Springs And The Geothermal Energy Potential Of Jammu & Kashmir State, N.W. Himalaya, India
[9] Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding On The Number Of Classes In Latent Class Analysis And Growth Mixture Modeling: A Monte Carlo Simulation Study. Structural Equation Modeling, 14, 535-569.
[10] Ravi Shankar Et Al., 1991. Geothermal Atlas Of India, GSI Spec Publ,
[11] WF Giggenbach Et Al, 1987. The Effects Of Hydrothermal Processes On The Chemistry Of Some Recent Volcanic Gas.