



Cluster Analysis on various samples of water pollution

Iqbal Kaur
Faculty of Mathematics, Govt P.G. College,
Karnal, Haryana-India

Abstract: The results of the verification of pollution are highly affected by the distribution of pH value, COD(Chemical Oxygen Demand), BOD(Biological Oxygen Demand), SS(Suspended Solids), and Oil and Grease, and depend mostly on the type of the industries. For pollution verification, in the present work the whole set of available measurements is divided into a set of industries, each of them displaying a similar precipitation pattern. Cluster analysis is adopted as an objective method to create groups of pollution factors displaying interrelated measurements.

Keywords – Pollution factors, cluster analysis, Principal component analysis.

I. Cluster Analysis or Classification Analysis

The classification is often referred to simply as discriminant analysis. In engineering and computer science, classification is usually called *pattern recognition*. Some writers use the term classification analysis to describe *cluster analysis*, in which the observations are clustered according to variable values rather than into predefined groups. In classification, a sampling unit (subject or object) whose group membership is unknown is assigned to a group on the basis of the vector of p measured values, y , associated with the unit. To classify the unit, we must have available a previously obtained sample of observation vectors from each group. Then one approach is to compare y with the mean vectors $y_1, y_2, \dots, y_k$ of the k samples and assign the unit to the group whose y_i is closest to y .

In Cluster analysis, the term *groups* may refer to either the k samples or the k populations from which they were taken. It should be clear from the context which of the two uses is intended in every case.

Here are some examples to illustrate the classification technique [3]:

1. A university admissions committee wants to classify applicants as likely to succeed or likely to fail. The variables available are high school grades in various subject areas, standardized test scores, rating of high school, number of advanced placement courses, etc.
2. A psychiatrist gives a battery of diagnostic tests in order to assign a patient to the appropriate mental illness category.
3. A college student takes aptitude and interest tests in order to determine which vocational area his or her profile best matches.
4. African, or "killer," bees cannot be distinguished visually from ordinary domestic honey bees. Ten variables based on chromatograph peaks can be used to readily identify them [5].
5. The Air Force wishes to classify each applicant into the training program where he or she has the most potential.
6. Twelve of the *Federalist Papers* were claimed by both Madison and Hamilton. Can we identify authorship by measuring frequencies of word usage [6].
7. Variables such as availability of fingerprints, availability of eye witnesses, and time until police arrive can be used to classify burglaries into solvable and unsolvable.
8. One approach to speech recognition by computer consists of an attempt to identify phonemes based on the energy levels in speech waves.
9. A number of variables are measured at five weather stations. Based on these variables, we wish to predict the ceiling at a particular airport in 2 hours. The ceiling categories are closed, low instrument, high instrument, low open, and high open [7].

II. Case Studies and Model Simulations

A. Data Collection

The pollution studies of Hisar City for pollution control has been carried out. The data was collected from the pollution control board, Hisar on the live variables, given below in table 1 for twenty different samples taken from the different places located in Hisar city of Haryana state of India.

X ₁	pH value
X ₂	COD(Chemical Oxygen Demand)
X ₃	BOD(Biological Oxygen Demand)
X ₄	SS(Suspended Solids)
X ₅	Oil and Grease

Table 1: Data of twenty industries about various pollution parameters.

Code	Type of Industry	X ₁ (pH)	X ₂ (COD) mg/l	X ₃ (BOD) mg/l	X ₄ (SS) mg/l	X ₅ (Oil and Grease) mg/l
A	Surgical unit	6.96	232.5	577.6	369	0
B	Card Board Industry	8.02	145	361.6	397	0
C	Textile unit	7.06	110	373.6	736	0
D	Cold Rolling	2.24	29	515.2	710	24
E	Steel Unit(Picling)	1.35	23.5	348	163	0
F	Steel Unit(Hot Rolling)	2.16	36	278.34	119	0
G	Textile unit	10.9	24	310.4	386	0
H	Surgical unit	8.71	119	492	249	0
I	Cold Rolling	1.18	27	294.4	510	12
J	Sewerage of Hisar	6.83	390	611.2	918	0
K	Textile unit	7.86	390	513.6	692	0
L	Steel Unit(Picling)	2.6	25	286.4	118	0
M	Distillery Unit	6.72	2750	10000	976	10
N	Steel Unit (Cold Rolling)	1.79	17	692	576	56
O	Surgical unit	10.13	220	395.2	396	0
P	Steel Unit(Picling)	6.34	22.5	497.6	170	0
Q	Agro Industry	8.82	92.5	296	190	0
R	Leather Unit	7.91	135	291.2	160	0
S	Milk Plant	6.9	320	579.2	463	0
T	Textile unit	10.91	48	286	396	0

B. Clustering Technique

In order to identify the most significant precipitation patterns in the wide number of available observation points (about 20 industries), suitable criteria to gather them around representative modes have been borrowed from clustering analysis techniques. Several hierarchical agglomerative clustering techniques have been tested in various meteorological applications. The average-linkage clustering algorithm [8] has been selected due to its good performances.

III. Results

A. Multivariate Analysis

Principal Component Analysis: the researchers first evaluated the mean of the five variables as given in the table 2.

Table 2: Shows the mean of the five variables

X ₁	6.2695
X ₂	257.8
X ₃	899.977
X ₄	464.7
X ₅	5.1

The computed Principal Component loadings are shown in table 3.

Table 3: Principal Component loadings

Code	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5
A	638.812	339.178	26.989	-2.191	5.469
B	409.332	376.572	-5.079	-5.074	6.497
C	431.126	707.124	-85.467	-14.394	4.524
D	544.477	659.909	-197.520	1.451	3.213
E	350.574	132.702	-87.413	-8.619	1.060
F	284.303	96.795	-52.001	-5.509	1.901
G	327.355	356.864	-106.096	-13.306	9.724

H	519.436	215.508	-43.994	-6.911	7.901
I	319.896	481.445	-116.280	-2.083	1.137
J	744.496	896.990	97.722	-1.587	2.993
K	637.540	682.077	151.893	4.639	4.683
L	289.088	93.863	-64.426	-6.538	2.385
M	10410.288	373.655	-51.895	-5.609	5.569
N	703.779	510.386	-239.259	31.248	7.404
O	461.488	380.893	58.122	-0.339	8.381
P	494.707	126.105	-127.368	-12.487	6.141
Q	320.360	171.766	-11.664	-3.799	8.100
R	325.274	147.218	34.015	0.198	7.154
S	668.950	441.775	97.959	2.569	4.803
T	345.349	965.445	-155.940	-22.792	7.580

The Principal Components are :

$$PC_{(1)} = 0.0X_1 + 0.265X_2 + 0.963X_3 + 0.058X_4 + 0.001X_5$$

$$PC_{(2)} = 0.003X_1 + 0.111X_2 - 0.09X_3 + 0.99X_4 + 0.009X_5$$

$$PC_{(3)} = 0.013X_1 + 0.955X_2 - 0.255X_3 - 0.129X_4 - 0.079X_5$$

$$PC_{(4)} = -0.123X_1 + 0.075X_2 - 0.020X_3 - 0.019X_4 + 0.989X_5$$

$$PC_{(5)} = 0.992X_1 - 0.003X_2 + 0.001X_3 - 0.003X_4 + 0.124X_5$$

Now from the table 5, it is clear that the first latent root correspond to 98.419% total variance of the given data. So it is enough to consider just first eigen vector which gives:

$$PC_{(1)} = 0.0X_1 + 0.265X_2 + 0.963X_3 + 0.058X_4 + 0.001X_5$$

B. Cluster Analysis or Classification Analysis

For cluster analysis of the given data(Table 1), compute the 20X20 matrix containing distances of different objects(A,B,---,T) from one another shown in table 4.

Table : 8 Distance Matrix

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A	0	234.728	437.39	402.72	372.64	436.68	374.33	186.04	377.47	572.13	365	436.89	9771.79	324.86	184.83	300.16	361.86	367.71	128.43	346.14
B	234.72	0	341.01	373.97	264.09	310.05	131.87	172.93	177.19	627.5	412.49	312.93	10001	400.93	82.21	291.6	223.4	247.43	286.93	123.01
C	437.39	341.01	0	166.99	580.08	628.69	365.93	501.26	253.8	409.85	316.12	304.44	9994.73	372.54	358.01	585.99	551.76	631.89	401.12	356.55
D	402.72	373.97	166.99	0	597.4	637.18	384.18	47.93	298.16	428.25	362.29	635.14	9670.31	224.46	387.44	540.87	568.41	603.75	387.79	390.05
E	372.64	264.09	580.08	597.4	0	83.33	226.34	193.14	351.33	879.57	402.22	76.31	10062.6	540.44	308.55	149.85	90.82	154.79	481.03	242.53
F	436.68	310.05	628.69	637.18	83.33	0	269.32	263.59	391.61	935.16	713.45	13.68	10129.71	619.24	352.56	225.55	92.67	1080.7	457.03	277.5
G	374.33	131.87	365.93	384.18	226.34	269.32	0	246.53	126.01	712.37	518.54	269.2	10083.18	430.1	231.79	285.87	208.13	407.19	407.2	35.65
H	186.04	172.93	501.26	470.93	193.14	263.59	246.53	0	340.34	731.58	519.76	261.35	9892.06	400.6	202.93	124.86	206.39	220.22	306.27	262.85
I	377.47	177.19	253.8	298.16	351.33	391.61	126.01	340.34	0	631.48	461.66	392.27	10091.11	405.55	370.32	522.72	326.94	366.55	411.51	117.24
J	572.13	627.5	409.85	428.25	879.57	935.16	712.37	731.58	631.48	0	246.17	937.41	9681.04	515.54	589.95	641.1	847.25	861.38	461.46	703.71
K	365	412.49	316.12	362.29	402.22	713.45	518.54	519.76	461.66	246.17	0	717.18	9779.68	433.11	361.3	638.59	622.78	630.48	248.28	506.35
L	436.89	312.93	304.44	635.14	76.31	13.68	269.2	261.35	392.27	937.41	717.18	0	10125.01	614.39	356.65	217.55	99.35	117.96	540.18	279.07
M	9771.79	10001	9994.73	9670.31	10062.6	10129.71	10083.18	9892.06	10091.11	9681.04	9779.68	10125.01	0	9709.289	9949.35	9918.9	10091.96	10087.86	9742.67	10099.46
N	324.86	400.93	372.54	224.46	544.44	619.24	430.1	405.55	515.54	433.11	614.39	9709.289	9709.289	0	406.08	453.66	560.97	592.27	347.07	448.79
O	184.83	82.21	358.01	387.44	308.55	352.56	231.79	202.93	370.32	589.95	361.3	356.65	9949.35	406.08	0	317.14	261.79	271.55	219.89	203.73
P	300.16	291.6	585.99	540.87	149.85	225.55	285.87	124.86	522.72	841.1	638.59	217.55	9918.9	453.66	317.14	0	214.35	209.66	425.45	310.67
Q	361.86	223.4	551.76	568.41	90.82	92.67	208.13	206.39	326.94	847.25	622.78	99.35	10091.96	560.97	261.79	214.35	0	52.25	454.41	210.99
R	367.71	247.43	631.89	603.75	154.79	108.07	407.19	220.22	366.55	861.38	630.48	117.96	10087.86	592.27	271.55	209.66	52.25	0	526.45	82.63
S	128.43	286.93	401.12	387.79	481.03	457.03	407.2	306.27	411.51	461.46	248.28	540.18	9742.67	347.07	219.89	425.45	454.41	526.45	0	405.53
T	346.14	123.01	356.55	390.05	242.53	277.5	35.65	262.85	117.24	703.71	506.35	279.07	10099.46	448.79	203.73	310.67	210.99	82.63	405.53	0

Table 4

Correlation analysis:

The correlation matrix is given below:

$$\begin{vmatrix} X_1 & X_2 & X_3 & X_4 & X_5 \\ X_1 & 1 & 0.1 & 0.33 & 0.19 & -0.49 \\ X_2 & 0.1 & 1 & 0.98 & 0.47 & 0.02 \\ X_3 & 0.33 & 0.98 & 1 & 0.43 & 0.11 \\ X_4 & 0.19 & 0.47 & 0.43 & 1 & 0.24 \\ X_5 & -0.49 & 0.02 & 0.11 & 0.24 & 1 \end{vmatrix}$$

From this matrix it is clear that maximum correlation is 0.98 between second and third variable i.e. second and third variable are most correlated. They depend upon each other. The regression line for these two variables is given as:

$$(X_2 - \bar{X}_2) = \frac{r_{0X_2}}{r_{0X_3}} (X_3 - \bar{X}_3)$$

$$X_2 - 0.273X_3 = 504.12$$

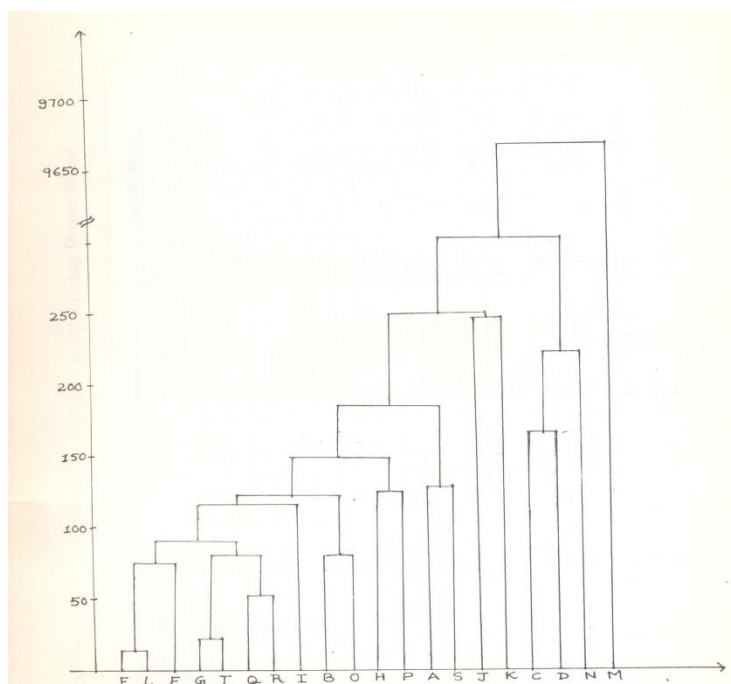


Figure1: Dendrogram for hierarchical cluster analysis using Single Linkage method.

Table 5: Details of clustering process for single linkage Distance method

Amalgamation order	No. of clusters	Distance	Cases	No. in cluster	Cluster description	Variable Means				
						X ₁	X ₂	X ₃	X ₄	X ₅
1	19	13.68	F,L	2	F,L	2.38	30.5	282.37	118.5	0
2	18	35.65	G,T	2	G,T	10.905	36	298.2	391	0
3	17	52.25	Q,R	2	Q,R	8.365	113.75	293.6	175	0
4	16	76.31		3	E,F,L	2.036	28.16	304.24	133.33	0
5	15	82.21	B,O	2	B,O	9.075	182.5	378.4	396.5	0
6	14	82.63		4	G,T,Q,R	9.635	74.875	295.9	283	0
7	13	90.82		7	E,F,L,G,T,Q,R	6.378	54.85	299.47	304.57	0
8	12	117.24		8	E,F,L,G,T,Q,R,I	5.73	51.37	298.84	330.25	1.5
9	11	123.01		10	E,F,L,G,T,Q,R,I,B,O	6.398	77.6	314.75	343.5	1.2
10	10	124.86	H,P	2	H,P	7.525	70.75	494.8	209.5	0
11	9	128.43	A,S	2	A,S	6.93	276.25	578.4	416	0
12	8	149.85		12	E,F,L,G,T,Q,R,I,B,O,H,P	6.58	76.46	344.76	321.16	1
13	7	166.29	C,D	2	C,D	4.65	69.5	444.4	723	12
14	6	184.83		14	E,F,L,G,T,Q,R,I,B,O,H,P,A,S	6.57	105	378.14	334.71	0.85
15	5	224.46		3	C,D,N	3.69	52	526.4	674	26.6
16	4	246.17	J,K	2	J,K	7.34	390	562.4	805	0
17	3	248.28		16	E,F,L,G,T,Q,R,I,B,O,H,P,A,S,J,K	6.72	140.62	401.17	393.5	0.75
18	2	253.8		19	E,F,L,G,T,Q,R,I,B,O,H,P,A,S,J,K,C,D,N	6.24	126.63	421.03	437.79	4.84
19	1	9681.04		20	ALL	6.2695	257.8	899.977	464.7	5.1

IV. CONCLUSIONS

Cluster analysis techniques have been applied to the aim of detecting any significant spatial variation in the pollution variants. The subsets detected by the clustering algorithm were remarkably persistent in the events, and pollution control appears to depend on the spatial variability represented by clusters. From the cluster analysis it is observed that the distillery unit is most polluted.

References:

- [1] Meredith A. Goldwasser and Aarrett A. Fitzmaurice, "Multivariate Linear Regression Analysis of Childhood Psychopathology using Multiple Informant Data", International Journal of Methods in Psychiatric Research.

- [2] Fitzmaurice, GM, Laird, NM, Zahner, GEP, Daskalakis, C. Bivariate logistic regression analysis of childhood psychopathology ratings using multiple informants. *American Journal of Epidemiology* 1995;142(11):1194-1203.
- [3] Alvin C. Rencher "Methods of Multivariate Analysis", Second Edition, Brigham Young University, a John Wiley & sons, Inc. Publication.
- [4] Rencher, A. C. (1998), *Multivariate Statistical Inference and Applications*, New York: Wiley.
- [5] Lavine, B., and Carlson, D. (1987), "European Bee or Africanized Bee?: Species Identification through Chemical Analysis," *Analytical Chemistry*, 59 (6), 468–470.
- [6] Mosteller, F., and Wallace, D. L. (1984), *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*, New York: Springer-Verlag.
- [7] Lachenbruch, P. A. (1975), *Discriminant Analysis*, New York: Hafner.
- [8] Bertò, A., A. Buzzi and D. Zardi, 2004: Back-tracking Water Vapour Contributing to a Precipitation Event over Trentino: a case study. *Meteorol. Z.*, **13**, 189-200.
- [9] Iqbal Kaur, Multivariate Analysis on various samples for water pollution, *American International Journal of Research in Science, Technology, Engineering & Mathematics*, 3(2), June-August, 2013, pp. 232-236.