A Multivariate Analytical Study on the Water of Han-River and the Streams flowing into Han-River Basin

Chul Lee* and Seungwon Kim

Department of Chemistry, Hanyang University, Seoul 133

Min-Young Kim

Seoul Metropolitan Government Institute of Health and Environment, Seoul 140. *Received June 23, 1987*

Pattern recognition techniques have been applied for the extraction of some regularities of water samples under a wide variety of locations related to Han-River. For that purpose, an eigenvector analysis has been applied for defining each class so as to use the class as a training set for class analogy model of SIMCA. The models thus obtained have been used for the allocation of test samples between groups.

Introduction

The search for regularities in empirical data has always been of major concern in environmental sciences. In many cases, only one or a few parameters that seem to be the best at first sight are used and many useful informations are therefore lost. The advances in the multivariate statistical techniques and the developments in the application of mathematical techniques for the analysis of many data have enabled the extraction of valuable informations.

It was generally found that most data obtained by analysis of environmental samples could not be approximated adequately with log-normal and with any other statistical tractable probability distribution.[1] Non-parametric procedures are therefore required. For that purpose, some pattern recognition techniques have been applied in this work as suggested by previous authors.[1] The method consisting of a combination of a conventional eigenvector analysis[2] and SIMCA (Statistical Isolinear Multicomponent Analysis)[3] has been found to extract the unique similarities of water samples under a wide variety of locations.

Data

The data for this study were taken from S.I.H.E.(Seoul Metropolitan Government Institute of Health and Environment) study one water pollution in the Han-River[4] and streams[5] which flow into the river. Sampling sites for 9 water reservoirs along Han-River and for 44 streams are shown in Table 1 and Figure 1. From each site, water samples were collected every month from January to December in 1984 and analyzed by means of standard methods.[6,7] The data from each site have been averaged for annual mean concentration (index i) made on sites (index k). The measured variables (i) are (1) water temperature(°C), (2) pH, (3) total alkalinity (mg/*l*), (4) total acidity(mg/*l*), (5) residue on evaporation (mg/*l*), (6) suspended solid(mg/*l*), (7) D.O.(mg/*l*), (8) B.O.D. (mg/*l*) (9) C.O.D.(mg/*l*) (10) soluble matter(mg/*l*), (11) sulfate(mg/*l*), (12) chloride(mg/*l*), (13) total sulfide(mg/*l*), (14) total hardness(mg/*l*), (15) ammonia nitrogen(mg/*l*), (16) nitrite nitrogen(mg/*l*), (17) nitrate nitrogen(mg/*l*), (18) relative safety(%), (19) perspective degree(cm), (20) total phosphate (mg/*l*), (21) A.B.S.(mg/*l*), (22) As(mg/*l*), (23) Cd(mg/*l*), (24) Pb(mg/*l*), (25) Cu(mg/*l*), (26) Fe(mg/*l*), (27) Mn(mg/*l*), (28) Zn(mg/*l*), (29) Ni(mg/*l*) and (30) coliform group(MPN/100 m*l*).

Table 1. Sampling sites of reservoirs

Classification	Sampling sites
Reservoirs	(1) Paldang (2) Dogso (3) Kuui (4) Dugdo (5) Bogwangdong (6) Noryangjin (7) Sonyu (8) Youngdeungpo (9) Gayang

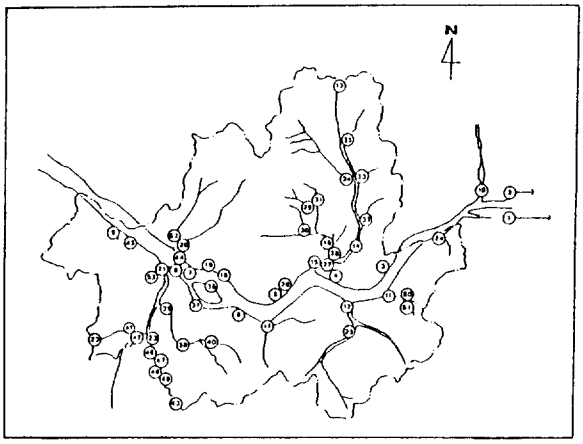

Figure 1. Water sampling sites for streams and water reservoirs.

In this work a datum may be missing when measured level is fallen below the detection limit. The missed data have been taken as zero for the multivariate analysis in this work.

Methods

The effect of very differing data ranges and variances of the various measured variables have been compensated by autoscaling the measured variables to produce "features" with means of zero and variances of unity. A separate preprocessing procedure was performed, *i.e.*, first transforming the measured values(x) to the ln(1.0 + x), then autoscaling as described above. The interdependence of 30 measured variables was evaluated through the inter-measurement correlation matrix.

The calculation of eigenvalues, eigenvectors and information preserved, as well as plots in the first-to-second and

first-to-third eigenvector planes have been made using conventional eigenvector analysis.[2] Members of each group have been decided on the basis of the spread of sample sites in the eigenvector plot and used as the training set.

Classification was further attempted by SIMCA using the training set as follows.[1,3]

The class q is defined by the parameters α, β, θ and ε in the principal component equation:

$$y_{ik}^{(q)} = \alpha_i^{(q)} + \sum_{a=1}^{A_q} \beta_{ia}^{(q)} \theta_{ak}^{(q)} + \varepsilon_{ik}^{(q)} \quad (1)$$

where y_{ik} is the feature of the i th constituent in the sample of the site k. A_q is the number of principal components. Before the model of eq. 1 can be used, e.g., for the classification of a new object, values of the parameters. $\alpha_i^{(q)}$, $\beta_{ia}^{(q)}$, $\theta_{ak}^{(q)}$ and σ_q^2 must be determined. The determination of $\beta_{ia}^{(q)}$ and $\theta_{ak}^{(q)}$ corresponds to the diagonalization of the matrices $Z^{(q)*} \cdot Z^{(q)*}$, where $Z^{(q)}$ denotes the matrix obtained from the feature matrix of q th reference data set after subtracting the average of each variable $\alpha_i^{(q)}$. Hence, values of the parameters β_{ia} and θ_{ak} have been obtained for each class. The deviation $\varepsilon_{ik}^{(q)}$ are then calculated by subtracting product terms of appropriate number A_q from the Z-value and yariances σ_q^2 are then estimated from these deviation as $S_0^{(q)2}$:

$$S_0^{(q)2} = \sum_{k}^{n_q} \sum_{i}^{M} (\varepsilon_{ik}^{(q)})^2 / [(n_q - A_q - 1)(M - A_q)] \quad (2)$$

where $i = 1,2, \ldots, M$ (M = number of variables), $a = 1,2, \ldots, A_q$ (A_q = number of the product terms in the model of eq. 1) and $k = 1,2, \ldots, n_q$ (n_q = number of objects in q th reference set). Thus, for the class q the model is calibrated by means of the data in the reference sets. The calibrated models can then be used to determine the classification of new objects as follows.

For the estimation of product terms A_q, the sum of the squares of the deviation Δ_A for each A-value for each object is calculated from each deviation ε_{ik}. The sum D_A is formed by adding the corresponding values Δ_A. These D_A-values are a measure of how well the model predicts the behavior of the reference-set for each value of A. By making F-tests on $(D_{A-1}-D_A)/(M-A_q)$ vs $D_A/[(n_q-A_q-1)(M-A_q)]$ one can determine whether the last product term (number A) is significant or not. The used critical F-value (p = 0.05) corresponds the number of degrees of freedom ($(M-A_q)$ vs $(n_q-A_q-1)(M-A_q)$).

The observed features of the object p, say y_{ip}, have been fitted to the model of eq. 1 with the same numbers of the product terms and with the same values of the parameters $\alpha_i^{(q)}$ and $\beta_{ia}^{(q)}$ as were obtained in the calibration of the model, using eq. 3.

$$y_{ip} - \alpha_i^{(q)} = Z_{ip} = \sum_{a=1}^{A_q} C_{ap} \beta_{ia}^{(p)} + \varepsilon_{ip}^{(q)} \quad (3)$$

For that purpose the parameter C_{ap} can be obtained by minimizing $\varepsilon_{ip}^{(q)}$, using β_{ia} values obtained in the calibration of model. The variance $S_q^{(q)2}$ of deviation (ε_{ip}) thus obtained then indicates how well the object p fits class q.

$$S_p^{(q)2} = \sum_{i=1}^{M} (\varepsilon_{ip}^{(q)})^2 / (M - A_q) \quad (4)$$

Result

A preliminary plot in the first-to-second eigenvector plane showed that the use of the ln(1.0 + x) transform and autoscaling gave somewhat greater separation between categories than simple autoscaling. This is reasonable, considering long

Table 2. Correlation Matrix Using Features of ln(1.0 + x). See Text for Concentration Variable Numbering(i)

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	1.0																													
2	0.17	1.0																												
3	0.43	0.05	1.0																											
4	0.41	-0.15	0.76	1.0																										
5	0.50	-0.06	0.72	0.80	1.0																									
6	0.24	-0.16	0.60	0.74	0.73	1.0																								
7	-0.25	0.06	-0.57	-0.65	-0.51	-0.52	1.0																							
8	0.43	-0.13	0.76	0.93	0.84	0.78	-0.64	1.0																						
9	0.45	-0.16	0.75	0.93	0.89	0.79	-0.63	0.98	1.0																					
10	0.62	0.10	0.71	0.75	0.92	0.63	-0.51	0.82	0.85	1.0																				
11	0.54	0.09	0.60	0.56	0.84	0.55	-0.42	0.60	0.67	0.84	1.0																			
12	0.65	0.14	0.74	0.76	0.91	0.66	-0.52	0.80	0.83	0.93	0.87	1.0																		
13	0.28	-0.12	0.38	0.45	0.44	0.34	-0.45	0.39	0.44	0.40	0.44	0.42	1.0																	
14	0.52	0.02	0.72	0.74	0.89	0.66	-0.49	0.82	0.83	0.86	0.78	0.87	0.45	1.0																
15	0.37	-0.04	0.78	0.90	0.80	0.68	-0.69	0.93	0.91	0.80	0.56	0.77	0.36	0.79	1.0															
16	-0.03	0.15	-0.33	-0.42	-0.30	-0.31	0.36	-0.44	-0.41	-0.32	-0.28	-0.30	-0.19	-0.31	-0.47	1.0														
17	0.14	0.05	-0.13	-0.11	0.09	0.07	0.22	-0.18	-0.10	-0.02	0.12	0.10	0.06	0.03	-0.28	0.23	1.0													
18	-0.40	0.05	-0.71	-0.85	-0.70	-0.69	0.67	-0.90	-0.87	-0.73	-0.49	-0.68	-0.33	-0.70	-0.91	0.43	0.38	1.0												
19	-0.25	0.19	-0.46	-0.60	-0.56	-0.57	0.49	-0.63	-0.65	-0.49	-0.39	-0.49	-0.30	-0.47	-0.55	0.27	0.02	0.57	1.0											
20	0.45	0.05	0.71	0.90	0.77	0.73	-0.58	0.89	0.89	0.76	0.54	0.73	0.49	0.71	0.89	-0.41	-0.23	-0.89	-0.54	1.0										
21	0.34	0.01	0.76	0.86	0.72	0.70	-0.58	0.88	0.84	0.69	0.52	0.68	0.38	0.69	0.87	-0.36	-0.20	-0.85	-0.61	0.86	1.0									
22	0.23	-0.14	0.07	0.23	0.29	0.21	-0.05	0.27	0.29	0.22	0.08	0.20	0.10	0.25	0.17	0.20	0.08	-0.23	-0.33	0.27	0.24	1.0								
23	0.51	0.15	0.67	0.69	0.72	0.67	-0.38	0.75	0.73	0.73	0.62	0.77	0.34	0.70	0.69	-0.27	0.01	-0.68	-0.44	0.71	0.73	0.27	1.0							
24	0.49	-0.21	0.42	0.46	0.64	0.44	-0.26	0.40	0.50	0.54	0.64	0.60	0.37	0.51	0.31	-0.11	0.28	-0.30	-0.39	0.40	0.33	0.31	0.54	1.0						
25	0.22	-0.20	0.34	0.36	0.63	0.38	-0.08	0.31	0.41	0.39	0.51	0.46	0.31	0.49	0.28	-0.02	0.47	0.12	-0.32	0.28	0.31	0.26	0.31	0.69	1.0					
26	0.23	-0.18	0.63	0.67	0.76	0.62	-0.53	0.63	0.67	0.59	0.61	0.65	0.36	0.69	0.63	0.25	0.22	0.48	-0.56	0.54	0.59	0.19	0.42	0.53	0.67	1.0				
27	0.13	-0.02	0.44	0.29	0.45	0.23	-0.25	0.31	0.33	0.48	0.43	0.45	-0.06	0.53	0.41	-0.16	-0.05	-0.30	-0.15	0.16	0.18	0.05	0.25	0.13	0.13	0.36	1.0			
28	0.27	0.12	0.53	0.49	0.66	0.48	-0.43	0.51	0.57	0.55	0.60	0.60	0.39	0.61	0.48	0.10	0.32	-0.34	-0.47	0.44	0.54	0.23	0.53	0.55	0.70	0.69	0.18	1.0		
29	0.62	0.27	0.25	0.27	0.54	0.33	-0.05	0.29	0.36	0.57	0.58	0.64	0.07	0.39	0.20	-0.01	0.54	-0.16	-0.29	0.27	0.23	0.24	0.51	0.58	0.52	0.35	0.06	0.55	1.0	
30	0.31	-0.05	0.67	0.81	0.73	0.72	-0.66	0.84	0.81	0.70	0.52	0.74	0.30	0.73	0.86	-0.44	0.19	0.76	-0.50	0.78	0.69	0.13	0.64	0.31	0.20	0.59	0.37	0.41	0.23	1.0

Table 3. Eigenvalues and Their Contribution(%)

Component	1	2	3	4	5	6	7
Eigenvalue	15.82	3.275	1.806	1.394	1.155	0.9876	0.8732
Contribution(%)	52.7	10.9	6.02	4.65	3.85	3.29	2.91
Cumulative(%)	52.7	63.6	69.6	74.3	78.2	81.5	84.4

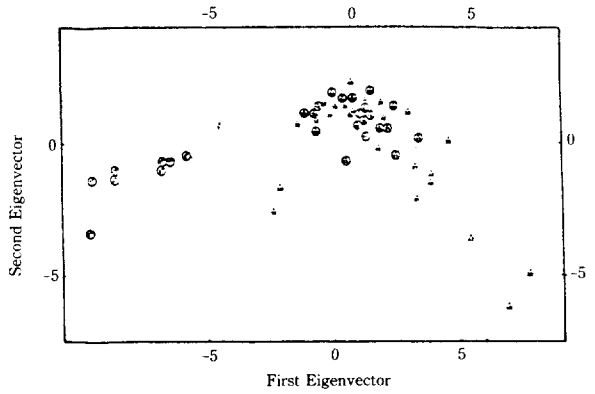

Figure 2. First-to-second eigenvector plot for the samples. Samples marked by asterisk are classified as belonging to class 1(water reservoirs), those marked by solid dot to class 2(stream-waters). The circled samples make up the training set for class 1 and class 2 respectively in SIMCA.

ranges of variables. The use of the ln(1.0 + x) transform can also make small difference less important while preserving major difference. Therefore, the features obtained by ln (1.0 + x) transform and autoscaling have been used for preprocessing data in this work.

As shown in Table 2, significant positive correlations over the entire data in the range 0.4-0.95 have been observed for the variables such as total alkalinity, total acidity, residue on evaporation, suspended solid, B.O.D., C.O.D., soluble matter, sulfate, chloride, total hardness, ammonia nitrogen, total phosphate, A.B.S., Cd, coliform group and relative safety.

The results on eigenvector analysis are summarized in the first-to-second eigenvector plot of Figure 2. The plot indicates that the sample points form two distinct categories, *i.e.*, one consists of nine water reservoirs and the other various stream waters with some outliers.

The eigenvalues, percentage of total variance within each eigenvector and cumulative percentage of variance are given in Table 3. The eigenvector coefficients corresponding to each eigenvalue have been calculated and are given in Table 4. The communalities of the variable in each of the derived eigenvectors, *i.e.*, the fraction of total variance accounted for in each component, are given in Table 4 to show the importance of given variable in an eigenvector. The table also shows that much important variables for the first eigenvector agree with the variables which give significant positive correlations described above. The separation between two categories is found to be mainly contributed to these variables. Table 4 also shows that important variables for the second eigenvector are nitrite nitrogen, nitrate nitrogen, relative safety, Pb, Cu, Zn, and Ni. Among these, some variables like nitrite nitrogen, nitrate nitrogen and relative safety much contribute to make samples from reservoirs a separated class. The outliers in Fig. 2 consist of samples from Anyang-Chun stream and the peculiarity of these samples could be attributed to the variables which give significant positive correlation described above and some heavy metals such as Pb, Cu, Zn and Ni.

Table 4. Eigenvector Coefficients and Communalities

	Eigenvector coefficients		Communalities	
i	comp.1	comp. 2	comp.1	comp.2
1	0.314	-0.174	0.286	0.099
2	-0.011	-0.079	0.002	0.021
3	0.206	0.067	0.674	0.015
4	0.229	0.116	0.829	0.044
5	0.237	-0.100	0.888	0.033
6	0.200	0.029	0.632	0.003
7	-0.165	-0.182	0.432	0.108
8	0.235	0.135	0.877	0.059
9	0.241	0.074	0.916	0.018
10	0.227	-0.055	0.815	0.010
11	0.196	-0.176	0.605	0.101
12	0.230	-0.106	0.840	0.037
13	0.122	-0.028	0.237	0.003
14	0.225	-0.047	0.798	0.007
15	0.228	0.186	0.821	0.114
16	-0.101	-0.212	0.161	0.147
17	-0.006	-0.440	0.001	0.634
18	-0.212	-0.230	0.710	0.174
19	-0.163	-0.022	0.420	0.002
20	0.221	0.142	0.774	0.066
21	0.215	0.134	0.729	0.059
22	0.070	-0.116	0.077	0.044
23	0.201	-0.039	0.639	0.005
24	0.150	-0.290	0.356	0.276
25	0.124	-0.349	0.245	0.399
26	0.189	-0.111	0.565	0.040
27	0.099	0.021	0.156	0.001
28	0.166	-0.236	0.437	0.183
29	0.118	-0.387	0.221	0.491
30	0.208	0.160	0.682	0.084

The plot of Figure 2 has only used the cumulative variance of 63.6%. This implies that 36.4% of the total variance is excluded, and obviously, some information about the grouping must be considered as lost by the projection of sites from thirty-dimensional observation space into the two-dimensional eigenvector plot of Figure 2. Therefore, SIMCA was applied in order to extract the ultimate of information from the data as follows.

The circled sampling sites of Figure 2 were chosen defining each class in the class analogy model of SIMCA. Values of the parameters $\alpha_i^{(q)}$, $\beta_{ia}^{(q)}$, $\theta_{ak}^{(q)}$ and σ_q^2 for q = 1.2 of eq. 1 and

Table 5(a). Parameter for Models of SIMCA. Class 1 for Water Reservoirs, Class 2 for Stream Waters, Concentration Variable Numbering(i) Same as in Table 2. β-values Normalized so that $\sum_{i=1}^{30} \beta_{ia}^2 = 1$

	Class 1(q = 1)				Class 2(q = 2)			
i	$\alpha_i^{(q)}$	$\beta_{i1}^{(q)}$	$\beta_{i2}^{(q)}$	$\beta_{i3}^{(q)}$	$\alpha_i^{(q)}$	$\beta_{i1}^{(q)}$	$\beta_{i2}^{(q)}$	$\beta_{i3}^{(q)}$
1	0	-0.1654	-0.3270	0.1815	0	-0.08611	0.1109	-0.1931
2	0	0.1650	-0.1081	0.06782	0	0.1329	0.1186	0.0006151
3	0	0.2261	-0.04115	0.09491	0	-0.05848	0.1418	-0.07143
4	0	0.2332	-0.04552	0.04708	0	-0.2568	0.1480	-0.1617
5	0	0.2155	-0.02790	0.1470	0	0.1423	0.3265	0.03303
6	0	0.2287	-0.04008	0.1196	0	0.2342	0.09100	0.2946
7	0	-0.2247	0.1573	-0.06145	0	0.1789	-0.1792	-0.01086
8	0	0.2384	0.01055	0.03791	0	0.2831	0.1147	0.03631
9	0	0.2389	0.02273	0.02683	0	-0.2034	0.2546	0.04985
10	0	0.03530	0.3572	0.4279	0	0.1875	0.2947	0.1442
11	0	-0.005554	0.08814	-0.5685	0	0.1272	0.3085	-0.04228
12	0	0.2134	0.01957	-0.1356	0	0.1109	0.3435	-0.1298
13	0	-0.07282	0.1485	-0.1651	0	-0.08683	0.2279	-0.2026
14	0	0.1948	-0.2489	0.01401	0	0.1127	0.3224	-0.005863
15	0	0.2369	0.05964	0.04513	0	-0.2345	0.07982	-0.1991
16	0	-0.06983	-0.1257	0.2514	0	0.1614	0.1130	0.2989
17	0	0.2002	0.08885	0.2857	0	-0.02696	0.1033	0.3210
18	0	-0.1614	0.3311	0.2124	0	0.3063	0.1131	0.1019
19	0	-0.2339	-0.005506	0.0005117	0	0.3446	-0.008177	-0.07532
20	0	0.2363	0.05300	0.04809	0	-0.2483	-0.04962	0.2400
21	0	0.2025	-0.2733	-0.02298	0	-0.2440	0.03791	0.1685
22	0	-1655	-0.1432	0.2028	0	-0.1312	-0.1705	0.2044
23	0	0.000	0.000	0.000	0	-0.09884	0.08441	-0.05060
24	0	0.0420	-0.2233	0.1944	0	-0.07049	0.1103	-0.1523
25	0	0.1266	0.4123	0.02581	0	-0.1058	0.1455	0.2820
26	0	0.2247	-0.1252	0.07048	0	0.1345	0.1673	0.2978
27	0	0.1810	0.2583	0.02895	0	0.2597	0.1563	0.1181
28	0	0.1758	0.2297	-0.1735	0	-0.1501	0.244	0.2311
29	0	-0.001368	-0.1038	-0.1961	0	-0.07387	0.002369	0.1842
30	0	0.2207	-0.1713	-0.09266	0	-0.1445	0.1499	-0.2967

Table 5(b). Parameter Values for θ_{ak}; Object Numbering(k) Same as in Figure 1. Class Numbering Same as in Table 5(a)

	Class 1(q = 1)				Class 2(q = 2)		
k	$\theta_{1k}^{(q)}$	$\theta_{2k}^{(q)}$	$\theta_{3k}^{(q)}$	k	$\theta_{1k}^{(q)}$	$\theta_{2k}^{(q)}$	$\theta_{3k}^{(q)}$
1	-5.002	0.2046	-0.2199	11	3.179	-0.8156	2.866
2	-4.489	1.045	1.989	13	-9.294	3.868	0.5240
3	-3.449	-1.182	-1.708	15	-0.3364	1.156	-0.6851
4	-2.171	0.3268	-1.112	16	1.541	-1.452	-0.2530
5	3.893	-3.165	-0.6011	18	-0.9905	-2.565	1.068
6	2.637	-0.6173	1.214	20	-2.323	-2.209	0.6493
7	3.578	0.4574	2.475	21	1.089	2.329	0.1486
8	5.003	2.931	-2.038	25	-2.389	-5.048	1.444
				28	0.8578	0.7477	-2.058
				29	-1.018	-2.525	0.3338
				31	0.3107	0.07100	-1.823
				33	2.782	-1.336	-0.7626
				34	-1.125	-0.2625	-2.633
				36	2.988	1.730	-1.189
				38	1.221	-1.129	-0.4085
				39	0.3417	-0.09806	0.1655
				41	1.889	3.323	4.874
				43	1.156	5.738	-0.3330
				44	-0.9095	-2.059	-0.3892
				53	1.109	0.5352	-1.538
S.D	4.170	1.766	1.702	S.D	2.721	2.544	4.874
Range	5.003	2.931	2.475	Range	3.179	5.738	4.874
	-5.002	-3.165	-2.038		-9.294	-5.048	-2.633

eq. 2 have been determined from the data of the reference set and are given in Table 5 and Table 6. For the estimation of the product number of similarity model, D_A values were calculated for each A value from the deviation ε_{ik} obtained by fitting, and necessary number of A value was found to be 3 for both classes by means of F-tests.

The values of S_p in eq. 4 for the object p in test set were calculated by using the similarity model. The distance $S_p^{(q)}$ corresponds to orthogonal distance between object p and the

Table 6. Distribution of Samples Between class 1 and class 2 Performed by SIMCA

Sample no.	Class given	Assignment calculated	Distance to nearest class, $S_p^{(q)}$	Distance to most distant class, $S_p^{(q)}$	$d_p^{(q)*}$
Training set for class 1					
1	1	1	0.4720(1)	3.220(2)	-
2	1	1	0.5043(1)	5.945(2)	-
3	1	1	0.4974(1)	3.313(2)	-
4	1	1	0.5227(1)	2.973(2)	-
5	1	1	0.4047(1)	2.219(2)	-
6	1	1	0.3584(1)	2.363(2)	-
7	1	1	0.3580(1)	2.369(2)	-
8	1	1	0.1801(1)	2.574(2)	-
Training set for class 2					
11	2	2	0.7532(2)	7.246(1)	-
13	2	2	0.4031(2)	4.203(1)	-
15	2	2	0.5948(2)	6.538(1)	-
16	2	2	0.6367(2)	7.287(1)	-
18	2	2	0.5490(2)	6.189(1)	-
20	2	2	0.6240(2)	5.610(1)	-
21	2	2	0.6339(2)	6.114(1)	-
25	2	2	0.6174(2)	5.632(1)	-
28	2	2	0.8602(2)	7.673(1)	-
29	2	2	1.064(2)	8.012(1)	-
31	2	2	0.6502(2)	6.110(1)	-
33	2	2	0.7099(2)	6.782(1)	-
34	2	2	0.6714(2)	7.064(1)	-
36	2	2	0.6354(2)	6.480(1)	-
38	2	2	0.5703(2)	7.083(1)	-
39	2	2	0.6665(2)	6.048(1)	-
41	2	2	0.6563(2)	9.415(1)	-
43	2	2	0.8255(2)	7.056(1)	-
44	2	2	0.6428(2)	6.223(1)	-
53	2	2	0.6118(2)	6.325(1)	-
Test set					
9	1	1	1.598(1)	1.935(2)	-
10	2	2	2.014(2)	10.95(1)	-
12	2	2	0.7804(2)	6.684(1)	-
14	2	2	0.6838(2)	6.799(1)	-
17	2	2	0.8100(2)	7.194(1)	-
19	2	2	0.6665(2)	5.769(1)	-
22	2	outlier	3.644(2)	32.03(1)	1079(C,C_2,C_3)
23	2	2	0.7683(2)	7.565(1)	-
24	2	2	0.9070(2)	6.525(1)	-
26	2	2	1.276(2)	5.067(1)	-
27	2	2	1.442(2)	10.95(1)	-
30	2	2	0.6314(2)	5.618(1)	-
32	2	2	0.8304(2)	7.247(1)	-
35	2	2	2.526(2)	7.284(1)	-
37	2	2	1.732(2)	6.756(1)	-
40	2	2	1.293(2)	6.664(1)	-
42	2	2	0.4782(2)	7.529(1)	-
45	2	2	1.342(2)	11.42(1)	-
46	2	2	1.417(2)	10.53(1)	-
47	2	outlier	3.228(2)	22.61(1)	3.976(C_2)
48	2	outlier	3.380(2)	25.60(1)	6.222(C_2)
49	2	2	1.389(2)	10.99(1)	-
50	2	2	1.777(2)	4.792(1)	-
51	2	2	1.884(2)	14.13(1)	-
52	2	2	1.385(2)	5.369(1)	-

standard deviation of training set of class 1(water reservoirs); $S_o^{(1)} = O_1 = 0.6022$. standard deviation of training set of class 2(stream waters); $S_o^{(2)} = O_2 = 0.7617$. critical F-value for class 1(water reservoirs); 1.72($A_q = 3$). critical F-value for class 2(stream waters); 1.70($A_q = 3$). *see reference 8.

class q. However, the actual distance $d_p^{(q)}$ between the object p and the class q has been determined by using $S_p^{(q)}$ as described elsewhere.[8] The object p is then assigned to the class q showing the smallest $d_p^{(q)}$ or to an outlier according to an F-test on $F = d_p^{(q)^2} / S_o^{(q)^2}$ using the number of degrees of freedom, *i.e.*, $(M\text{-}A_q)$ vs $(n_q\text{-}A_q\text{-}1)(M\text{-}A_q)$, as the critical F-value ($P = 0.05$). Classified results are given in Table 6.

An agreement has generally found between two methods, *i.e.*, and eigenvector plot in a two dimensional plane of Figure 2 and the results obtained by SIMCA which are given in Table 6. The peculiarity of some outliers, *i.e.*, samples from Anyang-Chun stream, have been found and will need further studies.

References

1. D. L. Duewer, B. R. Kowalski and T. F. Schatzki, *Anal. Chem.*, **47**, 1573 (1975).
2. C. Lee, O. C. Kwun, N. B. Kim and I. C. Lee, *Bulletin of Kor. Chem. Soc.*, **6**, 241 (1985).
3. S. Wold, *Pattern Recognition*, **8**, 127 (1976).
4. S. U. Han *et al.*, *Report of Seoul Metropolitan Government Institute of Health and Environment*, **20**(Cont'd) Issue), 369 (1984).
5. M. Y. Kim *et al.*, *ibid.*, P. 387.
6. JIS Standard Method for Industrial Waste Water.
7. A. P. H. A.: *Standard Method for the Examination of Water and Waste Water*, 14 th ed., Washington D.C., 1976.
8. S. Wold and M. Sjöström, Chemometrics, Theory and Application, ed., B. Kowalski, *Am. Chem. Soc. Symp. Ser.*, **52**, 243 (1977).