Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 2, pp. 349~357

Evaluation of Water Quality Using Multivariate Statistic Analysis with Optimal Scaling

Sang Soo Kim[1] · Hyun Guk Jin[2] · Jong Soo Park[3] · Jang Sik Cho[4]

Abstract

Principal component analysis(PCA) was carried out to evaluate the water quality with the monitering data collected from 1997 to 2003 along the coastal area of Ulsan, Korea. To enhance evaluation and to complement descriptive power of traditional PCA, optimal scaling was applied to transform the original data into optimally scaled data. Cluster analysis was also applied to classify the monitering stations according to their characteristics of water quality.

Keywords : Cluster analysis, MTV, Optimal scaling, Principal component analysis

1. 머리말

복잡하고 다양한 특성을 가지고 있는 해양의 수질을 가장 정확하게 파악하기 위해서는 조사 대상 해역의 해수 전체를 대상으로 수질상태를 조사하여야 하지만 시간적, 경제적인 제약으로 인하여 현실적으로 불가능하다. 따라서 해양 수질조사는 정해진 해역에서 일정 수의 조사정점을 지정해 놓고 주기적으로 이루어지는 것이 일반적이다.

한편 통계적 측면에서도 모집단의 통계치를 보다 정확하게 파악하기 위해서는 가능한 많은 수의 표본이 추출되어야 하는 것처럼, 보다 정확하게 해양 수질을 파악하기 위해서는 가능한 많은 수의 조사정점에서 수질조사가 이루어져야 하며 해마다 조사

1) 제1저자 : 국립수산과학원, 부산시 기장군 기장읍 시랑리 408-1, 619-902
E-mail : kimss@nfrdi.re.kr
2) 국립수산과학원, 부산시 기장군 기장읍 시랑리 408-1, 619-902
3) 국립수산과학원, 부산시 기장군 기장읍 시랑리 408-1, 619-902
4) 교신저자 : 608-736, 부산시 남구 대연동 110-1번지, 경성대학교 정보통계학과 부교수
E-mail : jscho@ks.ac.kr

정점의 수가 확대되고 있는 추세이고 우리나라 또한 그러하다. 그리고 채집된 시료를 이용하여 많은 항목이 조사, 분석되고 있으며 이에 따라 해마다 방대한 양의 자료가 생산되고 있다.

그러나 복잡하고 다양한 요인들의 영향을 받는 해양의 특성상 대부분 측정된 변수들이 서로 복잡하게 상관되어 있어 직접적 또는 직관적으로 수질상태 또는 측정변수들 간의 구조를 해석하기 어려운 경우가 많으므로 측정변수간 인과, 상관관계를 쉽게 파악할 수 있는 해석도구가 필요하다.

다변량 통계분석기법 중 주성분 분석(principal component analysis)은 서로 복잡하게 상관되어 있는 다변량 자료의 공분산행렬 또는 상관행렬을 이용하여 자료에 내재된 정보의 손실을 최소화하면서 자료를 축약하거나 새로운 합성차원(주성분)을 탐색하여 저차원의 그래프 상에 타점함으로서 시각적으로 자료를 해석하고 특성별로 자료를 분류하는 기법이다.

그러나 이러한 장점에도 불구하고 전통적인 주성분 분석을 통해 추출된 주성분을 적절하게 해석할 수 없거나 적정 수의 주성분에 대한 누적기여율이 낮은 경우 등, 원자료의 특성을 적절히 해석할 수 없는 경우가 발생하는 것이 대표적인 문제점으로 지적되고 있다.

그러므로 본 연구에서는 측정자료의 최적변환 방법을 이용함으로서 전통적인 주성분 분석이 갖고 있는 해석상의 문제점을 부분적으로 해결하고자 하였다. 즉, 최적 척도화(optimal scaling)방법으로 측정된 원자료(raw data)를 변환함으로서 원자료의 특성을 최대한 반영하되 분석모형과의 관계가 최대가 되도록 하여 주성분 분석을 시도하였다. 또한 최적 척도화에 의한 주성분 분석결과를 바탕으로 유사한 수질특성을 지닌 조사정점들을 서로 묶어줌으로써, 다수의 수질조사 정점을 소수의 동질적인 조사정점으로 군집화하여 동일 군집 내에 속해 있는 수질조사 정점의 수질특성을 평가하는 군집분석을 실시하였다.

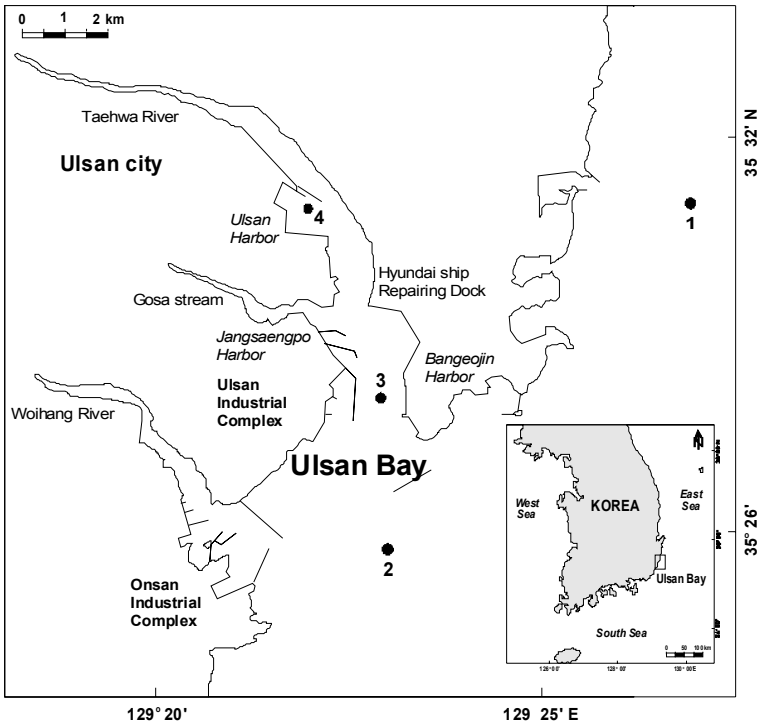
2. 조사방법

본 연구에 사용된 자료는 1997년부터 2003년까지 7년간 울산연안의 4개의 수질 조사정점에서 매년 정기적(2, 5, 8 및 11월)으로 표층해수를 분석한 조사결과로서(국립수산과학원, 1997~2003년) 수질 조사정점의 위치를 <그림 1>에 나타내었다.

분석항목 중 통계분석에 사용된 자료는 수온(Temp.), 염분(Sal.), 수소이온농도(pH), 용존산소(DO), 화학적 산소요구량(COD), 암모니아질소(NH4-N), 아질산질소(NO2-N), 질산질소(NO3-N), 총질소(T-N), 총인(T-P), 부유물질(SS) 및 투명도(Trans.)를 조사한 총 12개 항목의 자료였다.

계절별로 수질특성을 평가할 수 있는 측정항목을 판단하기 위하여 12개 항목으로 구성된 원자료를 이용하여 주성분분석을 수행하였으며, 고유치(eigenvalue)가 1.0이상인 성분을 대상으로 주성분을 결정하였다. 그리고 보다 저차원의 주성분 추출을 통해 수질특성을 평가하기 위하여 단조변환을 통해 최적인 값을 찾도록 하는 MTV(maximum total variance) 방법(Young, Takane and de Leeuw, 1978)을 사용하여 원자료를 변환하고, 변환된 자료를 이용하여 공분산행렬을 생성한 다음, 첫 번째부터 r번째 주성분까지 고유값(eigenvalue)의 합이 최대가 되도록 한 후 주성분 분석을 실시하였다. 또한 최적변환을 통해 추출된 주성분 분석결과를 바탕으로 계절별로 군

집분석을 실시하여 조사 정점들을 서로 유사한 수질특성을 갖는 그룹으로 군집화 하였다. 이 때 군집분석은 single-linkage agglomeration hierarchical 방법을 적용하였으며, 조사 정점들의 유사성을 거리로 환산하기 위해 조사 정점별 유클리디안 거리행렬을 사용하였으며 최적의 군집 수를 결정하기 위하여 'Semipartial R-square' 값을 사용하였다.



<그림 1> 울산연안 수질 조사정점 위치도

3. 결과 및 고찰

3.1 최적변환 자료를 이용한 주성분분석

12개 항목으로 구성된 원자료를 이용하여 주성분 분석을 수행한 결과는 <표 1>의 좌측에 나타내었다. 변환하지 않은 원자료를 이용한 주성분 분석에서 고유값이 1이상인 4개의 주성분(4차원)이 추출되었고 이들의 누적 기여율은 68.45%이였다. 이렇게 추출된 4가지 주성분 각각에 기여하고 있는 측정항목의 고유벡터의 절대값을 살펴보면, 첫 번째 주성분에서 고유벡터가 높은 측정항목은 NO2-N, NO3-N, T-N, T-P 및 SS이었으며 두 번째 주성분에서는 수온(Temp.)과 염분, 세 번째 주성분에서는 암모니아질소(NH4-N)와 투명도(Trans.), 네 번째 주성분은 pH, DO 및 COD의 고유벡터 값이 컸다. 첫 번째 주성분은 담수와 함께 해양으로 유입되는 영양염류를 설명할 수 있는

주성분으로 판단되고 두 번째 주성분은 계절적 변동을 나타내는 해양특성을 설명할 수 있는 주성분으로 볼 수 있었다.

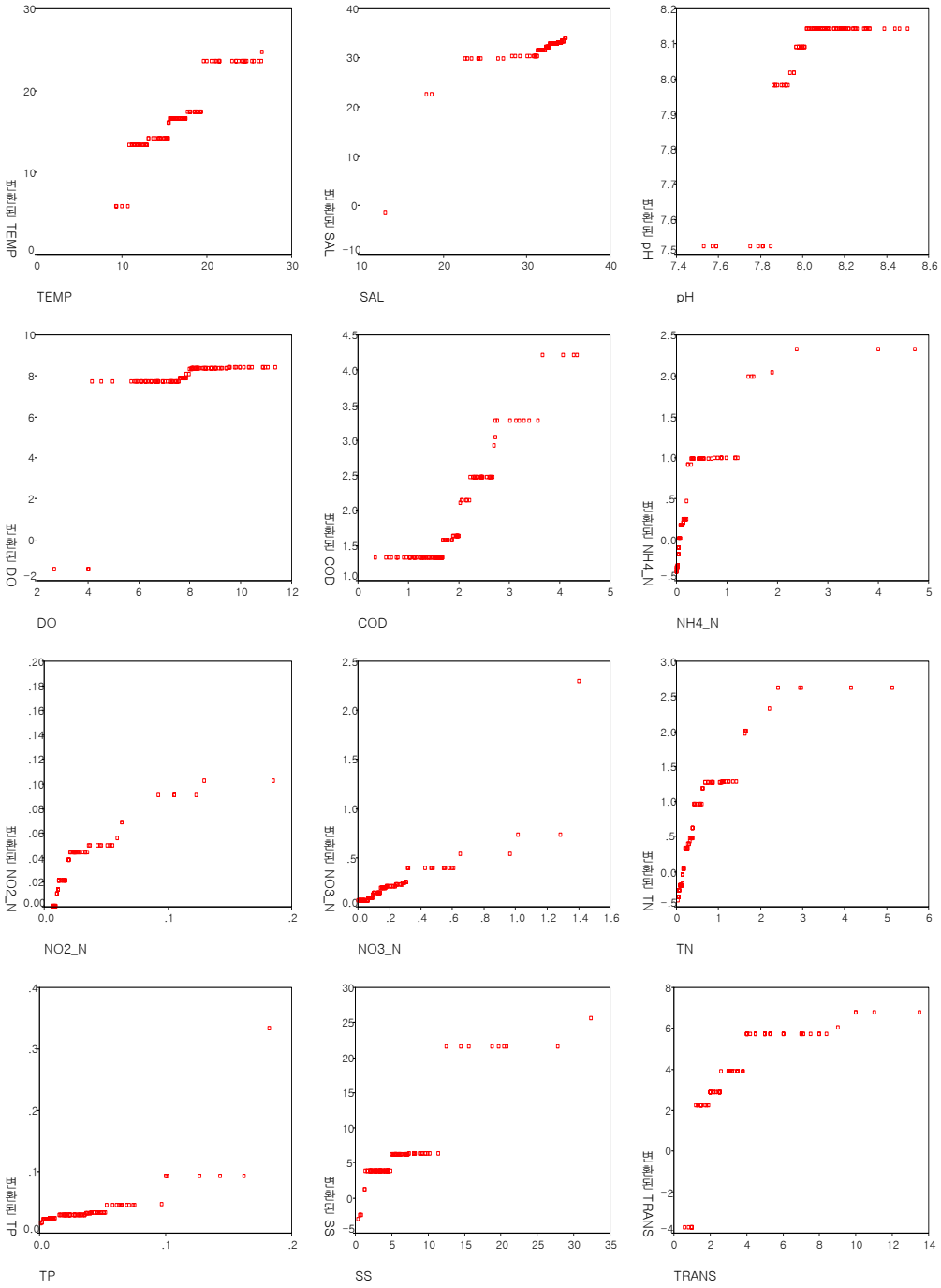
그러나 특히 세 번째 주성분에서 고유벡터 값이 높은 암모니아질소(NH4-N)와 투명도(Trans.)의 기여도를 설명할 수 있는 특성을 발견하기 어려우며, 네 번째 주성분에서 고유벡터 값이 큰 pH, DO 및 COD에 대한 설명도 해양에서 식물성 플랑크톤 양이 증가함에 따라 표층수의 용존산소 농도가 증가하기 때문으로 설명할 수 있겠으나 정확한 상관관계를 발견하기 어려운 점이 있다.

추출된 주성분의 수가 많을수록 원자료에 대한 설명력은 증가되지만 이상의 결과에서 알 수 있듯이 고유벡터의 값이 큰 측정항목들이 각 주성분에 대해 분산되어 있고 설명력 또한 분산되어 있으므로 주성분 분석의 가장 큰 장점인 자료 축약을 통한 확실한 결론 도출이 어려운 상황이다. 따라서 이러한 문제점을 부분적으로 해결하기 위하여 비슷한 누적기여율을 가지면서 더 낮은 차원의 주성분을 추출하여 원 자료를 설명하고자 한다. 즉, 최적화 척도법을 이용하여 원자료의 정보를 최대한 반영하되 측정자료와 분석모형간의 관계가 최대가 되도록 원자료를 최적변환하여 주성분 분석을 재시도하였다.

<표 1> 원자료와 최적변환자료를 이용한 주성분분석 결과

	원자료				최적 변환자료	
주성분	1주성분	2주성분	3주성분	4주성분	1주성분	2주성분
고유값	4.1853	1.6568	1.2127	1.1596	5.3795	2.3204
기여율	34.88	13.81	10.11	9.66	44.83	19.34
누적 기여율	34.88	48.68	58.79	68.45	44.83	64.17
측정항목	고유벡터				고유벡터	
	1주성분	2주성분	3주성분	4주성분	1주성분	2주성분
Temp.	-0.0085	-0.6438	0.2749	0.2193	0.0300	-0.4145
Sal.	-0.2779	0.4599	0.1765	-0.1430	-0.3279	0.3587
pH	-0.1412	-0.0720	0.0798	0.7117	-0.2235	0.0974
DO	-0.1645	0.3520	-0.1944	0.4298	-0.2409	0.3861
COD	0.2852	-0.0494	-0.0134	0.3881	0.2372	0.2843
NH4-N	0.3295	0.3435	0.4429	0.1436	0.2821	0.4034
NO2-N	0.3934	0.0551	-0.0274	-0.0029	0.3363	0.2676
NO3-N	0.3687	-0.0467	-0.3406	-0.0590	0.3790	-0.1604
T-N	0.4070	0.2831	0.2780	0.1057	0.3457	0.3416
T-P	0.3779	-0.1695	0.0467	-0.1653	0.3469	-0.2851
SS	0.2197	-0.0182	0.1107	-0.0559	0.2596	0.0152
Trans	-0.1953	-0.0989	0.6664	-0.1536	-0.2888	-0.0388

따라서 본 논문에서는 자료의 최적변환 방법의 일종인 단조변환을 통해 최적인 값을 찾도록 하는 소위 MTV(maximum total variance)방법을 사용하여 자료를 변환하였으며, 이 방법을 이용하여 최적변환시킨 결과와 최적변환 내용을 도식한 결과는 <그림 2>와 같다.

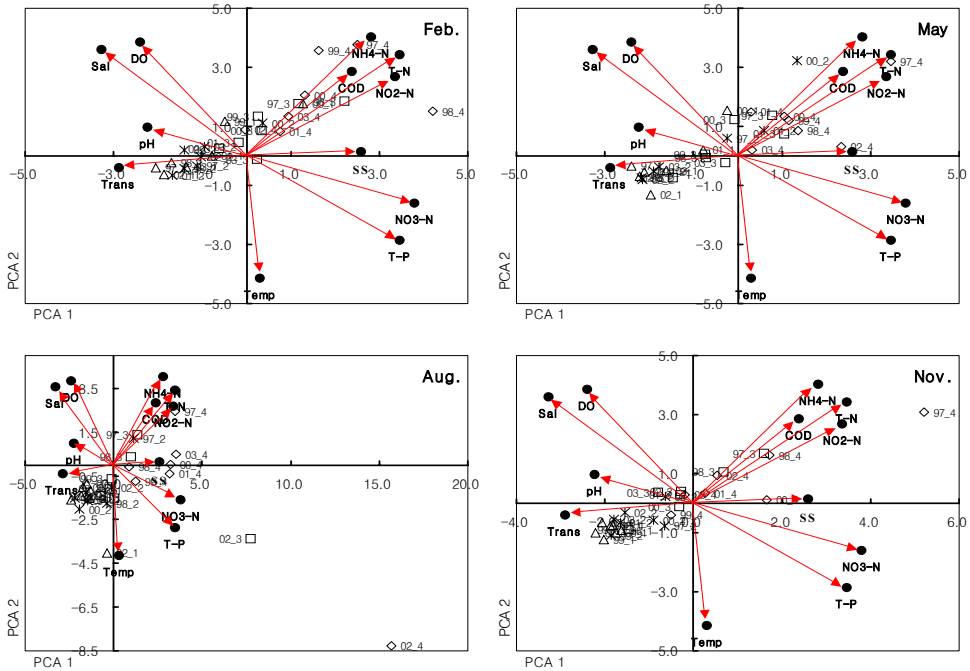


<그림 2> 최적변환식의 플롯

<그림 2>에서 나타난 바와 같이, 최적 척도화에 의해 최적변환된 자료의 특징을 살펴보면 다음과 같다.

1) 변수 pH는 S-곡선 형태로 변환된다. 이들은 중간수준에서 민감하게 변화하지만 중간수준의 앞부분과 뒷부분에서는 변동양이 거의 없다.
2) 변수 수온(Temp.), NH4-N, NO2-N, T-N, SS와 투명도(Trans.)의 최적변환은 오목함수의 형태이며, 제곱근 변환이나 로그변환 등과 같이 박스-칵스 변환에서 변환차수 k가 1 미만인 단조변환이 추천된다.
3) 변수 COD의 최적변환은 볼록함수의 형태이며, 제곱변환과 같이 박스-칵스 변환에서 변환차수 k가 1 보다 큰 단조변환이 추천된다.
4) 변수 SAL, DO, NO3_N과 TP는 대체적으로 선형의 형태이며, 박스-칵스 변환에서 변환 차수 k가 1인 단조변환이 추천된다.

<그림 2>에 나타낸 과정을 통하여 최적 변환된 자료를 이용한 주성분분석 결과는 <표 1>의 우측에 나타내었다. 원자료를 이용한 주성분 분석결과보다 작은 2개의 주성분이 추출되었고 누적 기여율도 64.17%인 것으로 나타나, 원 자료를 4개의 주성분으로 설명하는 경우와 비슷한 누적기여율을 가지고 있음을 알 수 있다. 최적변환 자료를 이용하여 추출된 2개의 주성분에 대한 bi-plot을 2차원 좌표 상에 계절별로 도식한 결과는 <그림 3>과 같다. 그림에서 도식된 레이블 중 '97_1'과 같은 레이블은 1997년에 조사정점 1에서 조사된 결과를 의미한다. 또한 그래프상에서 가시도를 높이기 위해 각 측정항목은 추출된 원값의 10배한 결과로 표시하였다.



<그림 3> 계절별 측정항목 및 조사정점의 고유벡터

먼저 <그림 3>에서 볼 수 있듯이 전통적인 주성분 분석에서 추출된 주성분에 대해 설명력을 가지고 있었던 pH, 수온(Temp.) 및 염분(Sal.)과 같은 측정항목은 최적변환을 통해 재시도된 주성분 분석에서는 설명력이 거의 없음을 알 수 있었으며 대부분 정점의 수질특성은 T-N, NO2-N 및 NH4-N과 같은 영양염류와 COD 및 투명도로 설명될 수 있음을 알 수 있었다.

그러므로 <그림 3>에 나타난 울산만의 수질특성을 영양염류, COD 및 투명도를 이용하여 설명해 보면, 우선 울산만의 경우 해역오염을 일으키는 주요 원인 중 하나인 도시하수의 대부분이 고사천이 유입되는 장생포항과 태화강을 통해 유입되므로(최민규 등, 2005) 두 유입원으로부터의 유달거리를 기준으로 판단해 볼 수 있을 것이다. 상대적으로 거리가 가까운 3 및 4번 정점은 육상으로부터 유입되는 유기오염의 지표가 될 수 있는 COD와 영양염류인 T-N, NO2-N 및 NH4-N의 영향을 강하게 받고 있음을 알 수 있으며 거리가 멀어질수록 이들의 영향은 적어지는 반면, 투명도의 영향이 커짐을 알 수 있다.

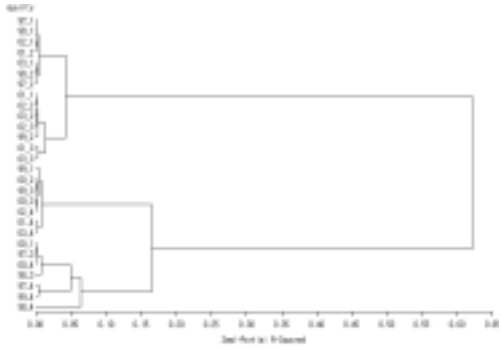
그리고 우리나라 강수량은 일반적으로 겨울철에 적다가 여름철로 갈수록 증가하는 특성을 보이며 육상으로부터 오염물질의 유입 절대량은 대체로 강수량이 증가할수록 비례하여 증가하는 것이 일반적인 현상이다(김학균, 2005). 그러므로 3 및 4번 정점은 5월과 8월로 갈수록 강수량이 증가함에 따라 하천을 통한 부유물질(SS) 유입량이 증가됨으로 인하여 SS의 영향을 받는 정점의 수가 연도별로 추가됨을 알 수 있다.

특히, 8월의 경우 연도에 따라 또 다른 오염물질인 NO3-N 및 T-P의 영향도 강하게 받고 있다는 사실도 알 수 있다. 그러나 11월의 패턴은 2월 결과와 유사한 것으로 판단할 수 있다.

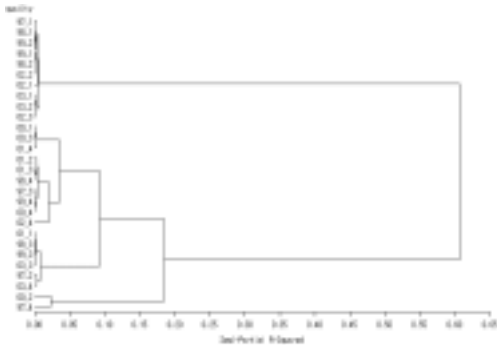
그러므로 울산만의 3번 및 4번 정점은 연중 지속적으로 COD와 영양염류인 T-N, NO2-N 및 NH4-N영향을 강하게 받고 있음을 알 수 있으며 또한 태화강을 통해 유입된 오염물질은 유달거리를 고려할 때 3번 정점 인근 해역에서부터 확산 또는 희석되기 시작함을 추론할 수 있었다.

3.2 최적변환 자료를 이용한 군집분석

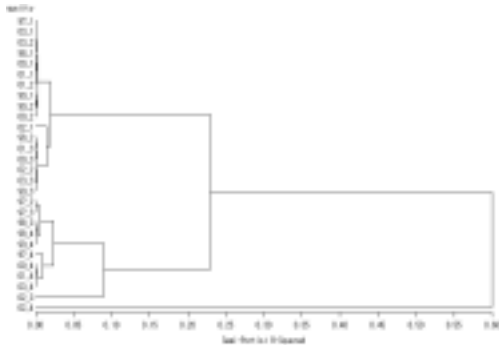
울산연안에 위치한 4개의 조사 정점에 대해서 최적변환 자료를 이용한 주성분분석 결과를 기초로 군집분석을 실시하여 각 조사월별로 조사정점의 군집화 되어가는 과정을 덴드로그램(dendrogram)으로 표시한 결과를 <그림 4> - <그림 7>에 나타내었다.



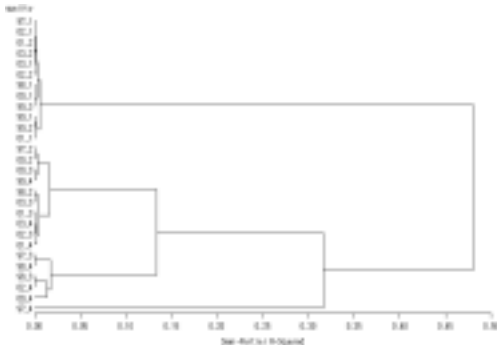
<그림 4> 2월 자료에 대한 군집분석 결과



<그림 5> 5월 자료에 대한 군집분석 결과



<그림 6> 8월 자료에 대한 군집분석 결과



<그림 7> 11월 자료에 대한 군집분석 결과

먼저 <그림 4>에 나타낸 2월 자료의 군집화 과정을 살펴보면, 최적의 군집 수는 3개인 것으로 나타났다. 3개로 분류된 군집 중 (99_1, 00_2, 99_3, 00_3, 02_4, 01_4, 03_4)로 구성되는 군집 1은 다른 군집에 비해서 군집 내의 유사성이 매우 높게 나타났으며 (97_1, 98_1, 02_1, 01_2, 03_1, 98_2, 97_2, 01_1, 02_2, 03_2, 02_3, 99_2, 01_3, 03_3)으로 구성되는 군집 2와 (00_1, 97_3, 00_4, 98_3, 97_4, 99_4, 98_4)로 구성되는 군집 3으로 나타났다.

그리고 <그림 5>에 나타낸 바와 같이 5월 자료의 최적 군집 수를 결정한 결과, 4개의 군집으로 분류되었으며 각각 분류된 군집 중 특히 (97_1, 98_1, 99_2, 99_1, 98_2, 02_2, 02_1, 03_1, 02_3)으로 구성되는 군집 1과 (01_1, 98_3, 99_3, 03_3, 97_2, 03_4)로 구성되는 군집 2는 각각 군집 내의 유사성 정도가 매우 높게 나타났으며, (00_1, 00_3, 01_4, 01_2, 01_3, 98_4, 97_3, 99_4, 00_4, 02_4)로 구성되는 군집 3과 (00_2, 97_4)로 구성되는 군집 4는 다른 군집에 비해서 이질적인 조사정점으로 나타났다.

<그림 6>의 8월 자료에 대한 군집화 과정에서도 최적의 군집 수는 4개인 것으로 나타났으며, 분류된 결과를 살펴보면 특히 (97_1, 03_1, 03_2, 98_1, 00_1, 01_1, 01_2, 99_1, 99_2, 00_2, 02_1, 98_2, 01_3, 00_3, 02_2, 03_3, 99_3)으로 구성되는 군집 1과 (97_2, 97_3, 98_3, 98_4, 99_4, 97_4, 00_4, 01_4, 03_4)으로 구성되는 군집 2는 군집 내의 유사성이 매우 높게 나타났으며, (02_3)와 (02_4)는 다른 조사정점과는 매우 다른 성질을 갖고 있음을 알 수 있었다.

11월 자료에 대한 군집화 결과는 <그림 7>에 나타내었으며, 같은 방식으로 최적의 군집 수를 결정한 결과 4개의 군집으로 분류되었다. 특히 (97_1, 02_1, 03_2, 03_1, 02_2, 98_1, 00_1, 99_3, 99_2, 01_1)로 구성되는 군집 1과 (97_2, 00_2, 00_3, 99_4, 98_2, 03_3, 01_3, 03_4, 02_3, 01_4)으로 구성되는 군집 2, (97_3, 98_4, 98_3, 02_4, 00_4)으로 구성되는 군집 3은 군집 내의 유사성이 매우 높으며, 이에 반해 (97_4)는 매우 이질적인 조사정점으로 나타났다.

이상의 계절별 각 조사정점의 군집분석 결과를 정리해 보면 우선 2월을 제외한 5월, 8월 및 11월 자료의 최적 군집 수는 4개인 것으로 나타났으며, 최적변환 과정을 통해 변환된 자료를 이용한 주성분 분석결과에서처럼 군집분석에서도 대체로 울산만으로 유입되는 가장 큰 유입원인 고사천과 태화강의 영향을 강하게 받는 3 및 4번 정점과 이와 반대인 1 및 2번 정점으로 크게 구별됨을 알 수 있었다. 군집 분류결과에서 1 및 2번 정점이 연도별, 계절별로 3 및 4번 정점과 유사한 군집으로 분류되기도

하고 이와 반대인 경우로도 군집이 분류되는 경우가 있었지만 이는 최적변환 방법을 이용한 원자료의 자료변환과정에서 원자료가 가진 정보가 최대한 반영되었지만 변환과정에서 일부의 정보는 소실되기 때문으로 생각된다.

그러나 전통적인 주성분 분석을 이용한 수질특성을 해석한 경우보다는 최적척도화법을 이용한 주성분 분석을 통해 수질특성을 평가하면 추출된 주성분에 대한 각 조사항목의 고유벡터 값을 최적화하여 집중시킴으로서 각 연도별로 조사정점의 수질특성에 대한 측정항목들의 영향과 계절별로 조사정점들이 어떻게 변화되는지를 보다 단순화시켜 보여줄 뿐만 아니라 자료의 해석도 용이함을 알 수 있었다.

참고문헌

1. 국립수산과학원 (1997, 1998, 1999, 2000, 2001, 2002, 2003) *한국해양환경조사연보.*
2. 허명회 (1994), *SAS 최적척도법*, 자유아카데미.
3. 최민규, 최희구, 김상수, 문효방 (2005) Fecal sterol을 이용한 울산만과 주변해역 퇴적물내 하수기인 유기물 평가, *한국환경과학회지*, 14, 23-32.
4. 김학균 (2005) *해양적조*, 다솜출판사.
5. Breiman, L. and Friedman, J.H. (1985) Estimating optimal transformations for multiple regression and correlation, *Journal of American Statistical Association*, 77, 580-619.
6. Davison, M.L. (1983) *Multidimensional scaling*, Wiley, New York.
7. Everitt, B.S and Dunn, G. (1991) *Applied Multivariate Data Analysis*, Edward Arnold, London.
8. Gabriel, K.R. (1971) The biplot graphics display of multivariate matrices with applications to principal component analysis, *Biometrika*, 58, 453-467.
9. Johnson, R.A. and Wichern, D.W. (1988) *Applied multivariate analysis*, Second Edition, Prentice-Hall, London.
10. Winsberg, S. and Ramsay, J.O. (1983) Monotone transformations for dimension reduction, *Psychometrika*, 48, 575-595.
11. Young, F.W. Takane, Y. and de Leeuw, J. (1978) The principal components of mixed measurement level multivariate data : An alternating least squares method with optimal scaling features, *Psychometrika*, 43, 279-281.

감사의 글 자료제공을 해 주신 국립수산과학원에 감사를 드립니다. 이 논문은 국립수산과학원 연구실적 "RP-2005-ME-016"호입니다.

[2005년 3월 접수, 2005년 5월 채택]