한국환경과학회지 제17권(제5호), 517~524, 2008
Journal of the Environmental Sciences

지역 · 시간별을 고려한 이차원 대기환경 군집 분석

위성승[*,**] · 김재훈[*] · 안치경[**] · 최병수[***] · 김대선[*]
[*]국립환경과학원 환경역학과, [**]성균관대학교 통계학과, [***]한성대학교 멀티미디어공학과
(2008년 1월 15일 접수; 2008년 3월 4일 채택)

Two Dimensional Cluster Analysis of Air Quality by Time and Area

Seongseung Wee[*,**], Jaehoon Kim[*], Chikyung Ahn[**],
Byongsu Choi[***] and Dae-seon Kim[*]

[*]*Environment Epidemiology Division, National Institute of Environmental Research, Incheon 404-708, Korea*
[**]*Department of Statistics, SungKyunKwan University, Seoul 110-745, Korea*
[***]*Department of Multimedia Engineering, HanSung University, Seoul 136-792, Korea*
(Manuscript received 15 January, 2008; accepted 4 March, 2008)

Abstract

The purpose of this study was to investigate the characteristics of air quality using data from which obtain local air quality monitoring system for cohort study in Chungju, Korea. We analyzed the concentration data of NO_2, SO_2, and PM_{10} in Chungju and industrial cities in 2006. We compared a industrial area with a cohort study area using by bicluster algorithm. In the case of SO_2, the rate of the cluster time was 10~60% and the cluster time number of two areas was similar. In the case of NO_2 and PM_{10}, the number of cluster time between a industrial area and cohort study area was clearly different.

Key Words : Bicluster, Cluster analysis, Cohort, Air pollution, Air quality monitoring system

1. 서 론

최근 우리나라 대규모 국가산업단지, 폐광 등 환경오염이 우려되는 지역의 주민들이 피부병, 천식 등 각종 환경성 질환을 호소하는 경우가 증가하고 있어, 환경오염과 지역주민의 건강영향에 대한 다양한 조사가 진행되고 있다[1]. 이들 환경오염 우려지역 주민에 대한 조사결과의 객관성을 높이기 위해서는 환경오염에 의한 건강 악영향이 배제된 지역을 대상으로 대조코호트 조사가 필요하다. 코호트 조사의 목적은 특정한 요인을 가진 인구집단을 선정하여 장기간에 걸친 관찰과 조사를 통해 해당 요인이 건강에 어떠한 영향을 미치는지를 확인 · 관찰하는데 있다. 본 연구에서는 국가산업단지 지역에 대한 대조코호트 조사지역을 선정하기 위해 ①내륙(해안)에 위치한 도시 ②국가산업단지가 인근에 위치해 있지 않은 도시 ③인구 15만명 이하 도시 ④연령 및 성별 인구 구성이 산업단지 지역과 비슷한 도시 ⑤대기오염자동측정망이 설치된 도시 등 5가지 기준을 바탕으로 후보도시들을 1차적으로 선정하였고, 관련 자료 검토를 통해 최종적으로 내륙지역

Corresponding Author : Jaehoon Kim, Environment Epidemiology Division, National Institute of Environmental Research, Incheon, 404-708, Korea
Phone: +82-32-560-7273
E-mail: clean@me.go.kr

인 충주와 해안지역인 강릉을 대조코호트 조사지역으로 각각 선정하였다. 본 연구에서는 이들 두 지역에 대해 본격적인 코호트조사에 앞서 대기오염물질 측정망 자료를 바탕으로 선행연구차원에서 지역과 시간을 고려한 군집분석을 통하여 지역간 대기오염 특징차이에 대해서 알아보고자 하였다.

본 연구에서는 산업단지 지역인 울산, 포항, 광양, 시화 · 반월과 대조지역인 충주, 강릉의 2006년 1년간의 대기오염자동측정망 자료 중 아황산가스(SO_2), 이산화질소(NO_2), 미세먼지(PM_{10}) 등 3개 항목을 이용하여 두 지역간에 얼마나 유사한 변동을 하는지에 대해 통계적으로 분석함으로써 두 지역간의 대기환경의 특징들을 살펴보았다. 기존 관련 연구에서는 일반적으로 각 지역별로 농도의 차이나 지역의 k-평균 군집분석 방법[2]을 적용시켜 대기오염물질 특징에 관한 연구가 주로 진행되어 왔으나, 이러한 농도 비교분석은 자료를 단순화시키는 면이 있고, k-평균 군집분석은 k를 얼마나 할 것인가의 전문가적 판단에 의해야 하며 재군집시에 똑같은 결과가 나오지 않는다는 단점을 가지고 있다. 또한 기존의 군집분석은 1개 요인만의 정보를 이용하여 군집을 하였기 때문에 시간의 장기적인 부분과 결측값에 대해서는 군집을 찾아내기 매우 난해하였다. 또한 단순히 지역간의 평균비교를 했을 때 분석결과는 지역간의 관측소의 개수를 무시한 단순한 평균농도차이를 표현한다. 이러한 단점을 극복하기 위해 본 연구에서는 바이클러스터 알고리즘을 이용하여 지역내 대기측정소 개개의 차이까지 고려하므로써 산업단지 지역내 어느 지역이 대조지역과 얼마만큼의 유사한 변동을 보이는지를 통계적 수치로 분석하는데 초점을 두었다.

본 연구의 결과는 산업단지지역과 대조지역의 대기환경특성을 비교하는데 수치적 비교자료가 될 뿐 아니라, 향후 신규 코호트 조사지역을 선정을 위한 방법으로 활용될 수 있을 것이다.

2. 재료 및 방법

2.1. 실험지역

우리나라 대규모 국가산업단지가 입지해 있는 울산, 포항, 광양만(여수시), 시화 · 반월과 대조지역인 충주 및 강릉시를 연구대상지역으로 하였다.

2.2. 실험방법

국가산업단지 및 대조지역의 주요 대기오염물질인 SO_2, NO_2, PM_{10} 3개 항목에 대해 기존 대기환경자동측정망[3]을 통해 측정된 2006년 1월부터 2006월 12월까지 자료를 이용하였다. SO_2, NO_2농도는 지역내 산업활동, 자동차운행, 석유, 석탄과 같은 화석연료 사용량에 밀접한 관련이 있으며 PM_{10}은 건강에 천식 및 호흡기, 심혈관계질환에 영향을 미치는 것으로 알려져 있다[4,5].

자료분석을 위해 통계학적 알고리즘인 행과 열을 고려한 바이클러스터 알고리즘을 이용하여 각 물질의 지역별 시간대별 농도에 대해서 변동을 분석하였고, 통계프로그램인 R.2.6.1 패키지를 이용하여 바이클러스터 알고리즘을 구현하였다. 바이클러스터 알고리즘은 행과 열을 이용한 이차원을 고려하는 알고리즘이며 Cheng and Church[6]의해 제안되었다. 바이클러스터 알고리즘은 구현하기 매우 복잡하며, 마이크로어래이 분야에서 gene과 array를 이용한 이차원의 자료를 군집하기 위한 방법으로 개발되었으나, 알고리즘이 난해하여 널리 활용하지 못하다가 최근에 유전체 연구에서 유의한 유전자를 찾기 위한 방법으로 매우 관심이 높은 알고리즘이다. 본 연구에서는 지역과 시간을 고려한 이차원을 고려해야 하기 때문에 바이클러스터 알고리즘을 적용하였다.

2.3. 통계적 방법

2.3.1. 군집분석과 바이클러스터 알고리즘 비교

Fig. 1과 같이 군집분석 방법은 열을 고려하지 않는 행적인 부분만 고려한 방법이며 Gene1, Gene4와 Gene9를 한 군집으로 보게 된다[7]. 반면에 바이클러스터링은 행과 열을 고려하기 때문에 행에서는 Gene1, Gene4 와 Gene9를 열에서 A, B, D, E, F, G, H의 Conditions를 같은 군집으로 하여 찾아낸다. 또한 녹색부분만 보았을 경우 Gene1, Gene4, Gene6, Gene7, Gene9와 B, E, F의 Conditions를 찾게 된다[3].

2.3.2. 바이클러스터 알고리즘 방법

$A=\{a_1, a_2, ..., a_N\}$ 을 지역집합, $T=\{t_1, t_2, ..., t_M\}$ 을 시간집합이라 하자. 지역들과 시간들로 이루어진 데이터 S 는 실수값들로 구성된 $N\times M$ 의 행렬로 보여 질 수 있다

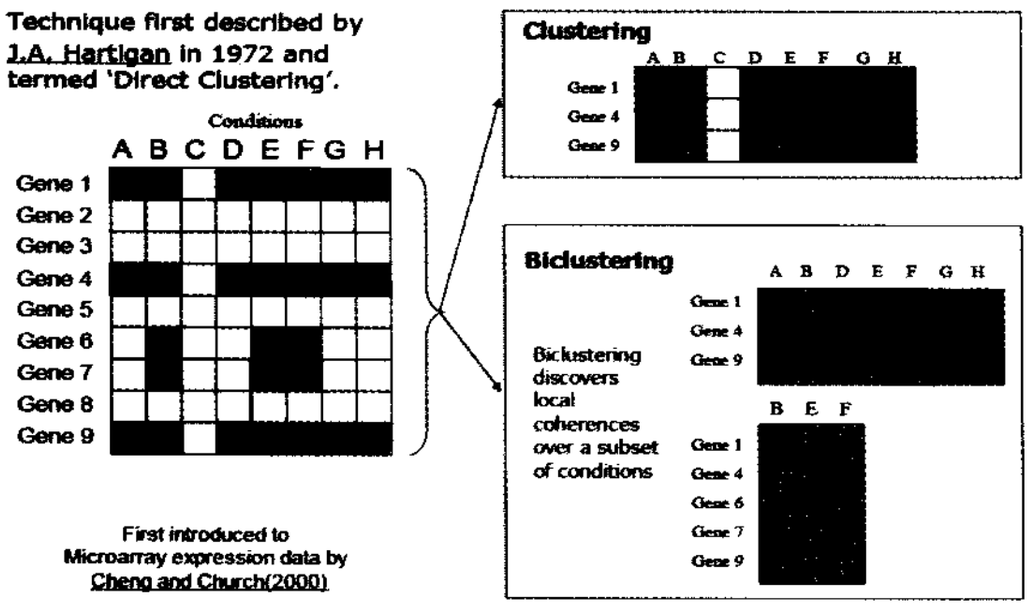


Fig. 1. Compare clustering and biclustering using gene expression data.

행렬에 있어서 각각의 엔트리 a_{ij} 는 특정시간 t_j 하에서 특정 지역 a_i의 농도 수준을 의미한다[8]. 바이클러스터는 전체 행렬에서 행이나 열의 응집성을 보이는 부행렬(Subset Matrix)를 말한다. 행인덱스의 집합을 I 이라고 하고, 열인덱스를 J 라고 할 때, 바이클러스터는 $|I| \leq |N|$ 과 $|J| \leq |M|$ 인 (I, J) 이리고 표기할 수 있다. 바이클러스터는 평균 최소잔차(Mean Squared Residue, MSR) 스코어를 최소화하는 바이클러스터를 찾는데 목적이 있다[9].

MSR Score 는

$$H(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} R(a_{ij})^2$$

로 정의 되며, 여기서 I 와 J 에 의해서 결정되는 바이클러스터의 각각 엔트리 a_{ij} 의 residue는 다음과 같이 정의되고

$$R(a_{ij}) = a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}$$

각 항목 a_{iJ} 은 행 i의 평균, a_{Ij} 은 열 j 의 평균, a_{IJ} 는 행렬 A의 평균이다.

H 값이 높으면 자료간의 유의성이 낮아지며, H 값이 낮으면 행렬들 간의 상관성이 의미가 있다고 볼 수 있다[10]. 본 연구에서 사용한 바이클러스터 알고리즘은 기존 연구[6]에서 검증된 바 있다.

3. 결 과

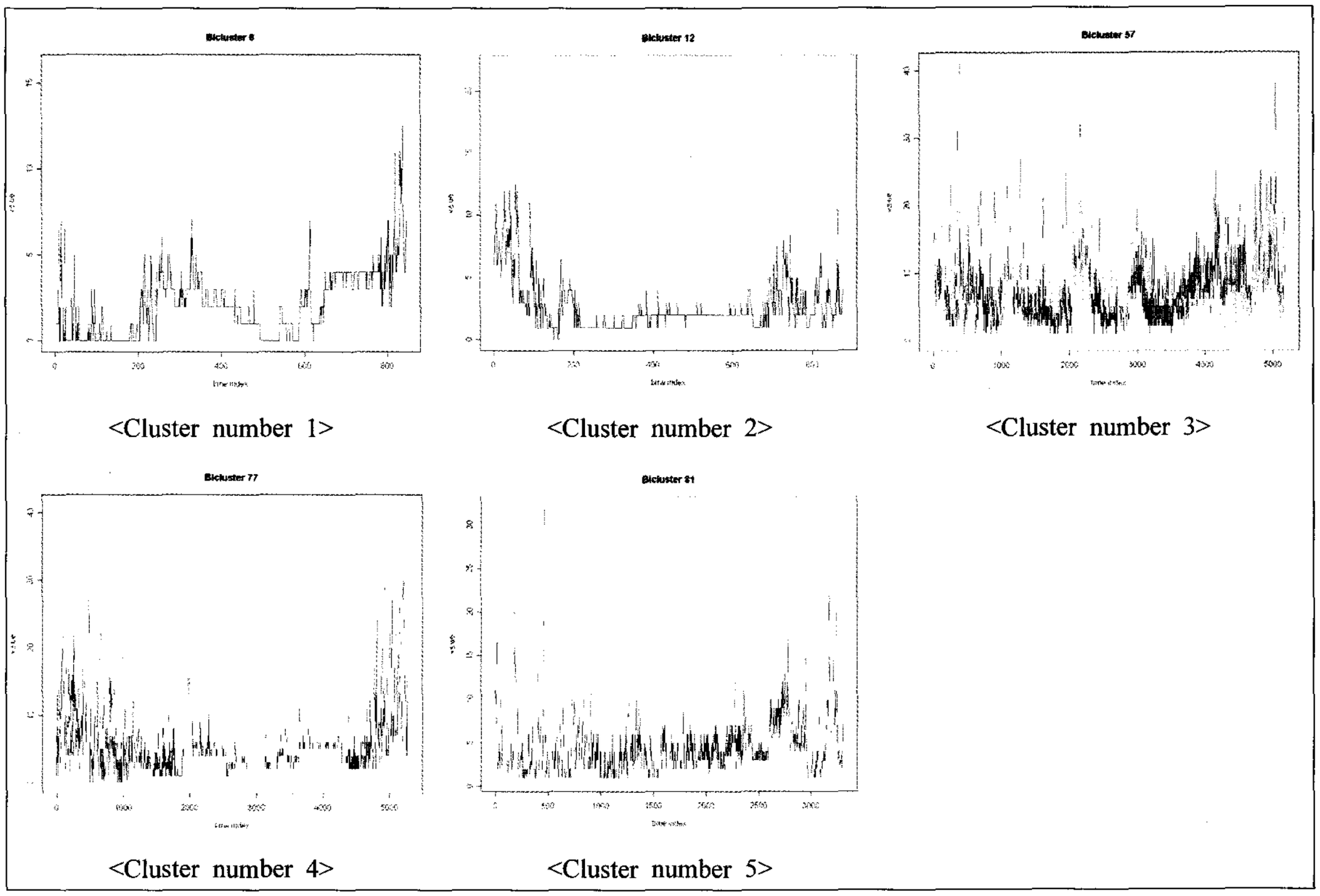
SO_2, NO_2, PM_{10} 3개 항목에 대해 충주 1개지역, 강릉 1개 지역, 울산 14개지역, 포항 4개지역, 광양만권(여수포함) 6개지역, 시흥 1개 지역에서 각각 측정된 자료를 분석하였다. 각 물질의 농도값의 단위는 SO_2, NO_2는 ppb, PM_{10}는 $\mu g/m^3$ 이다. 2006년 1월부터 2006년 12월까지 시간시점은 8760 이다. 이는 하루 24번 측정하고 365일 측정값임을 의미한다. 각 물질마다 100번의 군집을 실행하여 그 중에 가장 의미 있는 군집결과 5개를 보여주며 군집지역과 군집시간시점에 대한 빈도와 전체 시간시점에 대비 % 를 나타낸 것이다. 단 가급적 군집의 결과에서 산업단지 및 대조지역을 먼저 선정하였으며, 지역이 같은 도시의 군집은 본래의 연구목적에 벗어나므로 제외하였다.

3.1. SO_2 물질 분석 결과

대기중 SO_2에 대한 분석결과를 Fig. 2와 Table 1에 정리하였다. 1집단으로 울산시 덕신리와 충주시 문화동, 2집단으로 울산시 효문동과 강릉시 옥천동, 3집단은 광양만권인 여수시 광무동과 포항시 장흥동 4집단은 충주시 문화동과 시흥시 시화공단, 5집단은 울산시 신정동과 강릉시 옥천동이다. 대기중 SO_2의 군집분석 결과, 1군집인 울산시 덕신리와 충주시 문화동은 약 9.6%의 유사성를 보였으며, 2집단인 울산시 효문동과 강릉시 옥천동은 약 10%의 유사성, 3집단인 광양만권인 여수시 광무동과 포항시 장흥동은 약 58.94%로 유사성을 보였으며 4집단인 충주시 문화동과 시흥시 시화공단은 약 60.32%, 5집단인 울산시 신정동과 강릉시 옥천동은 약 37.79%의 유사성을 보였다. 분석결과 대조집단과 산업단지 집단간에 SO_2에 대해 최소 10%에서 최대 60%의 지역간의 동질적인 변동을 나타난 것으로 확인되었다.

3.2. NO_2 물질 분석 결과

대기중 NO_2 농도에 대한 분석결과를 Fig. 3과 Table 2에 정리하였다. 1집단으로 울산시 상남리, 포항시 죽도동, 포항시 대도동, 포항시 대송면, 충주시 문화동, 2집단으로 울산시 덕신리, 울산시 상남리, 포항시 죽도동, 포항시 대도동, 충주시 문화동, 3집단은 포항시 죽도동, 포항시 대도동, 포항시 대송면, 강릉시 옥천동 4집단은 포항시 장흥동, 포항시 죽도동, 포항시 대송면, 충주시 문화동, 5집단은 여수시 삼일동, 포항시 죽도동, 포항시 대송면, 강릉시 옥천동이다. 대기중 NO_2 농도에 대한 군집결과 1집단인



<Cluster number 1> <Cluster number 2> <Cluster number 3>

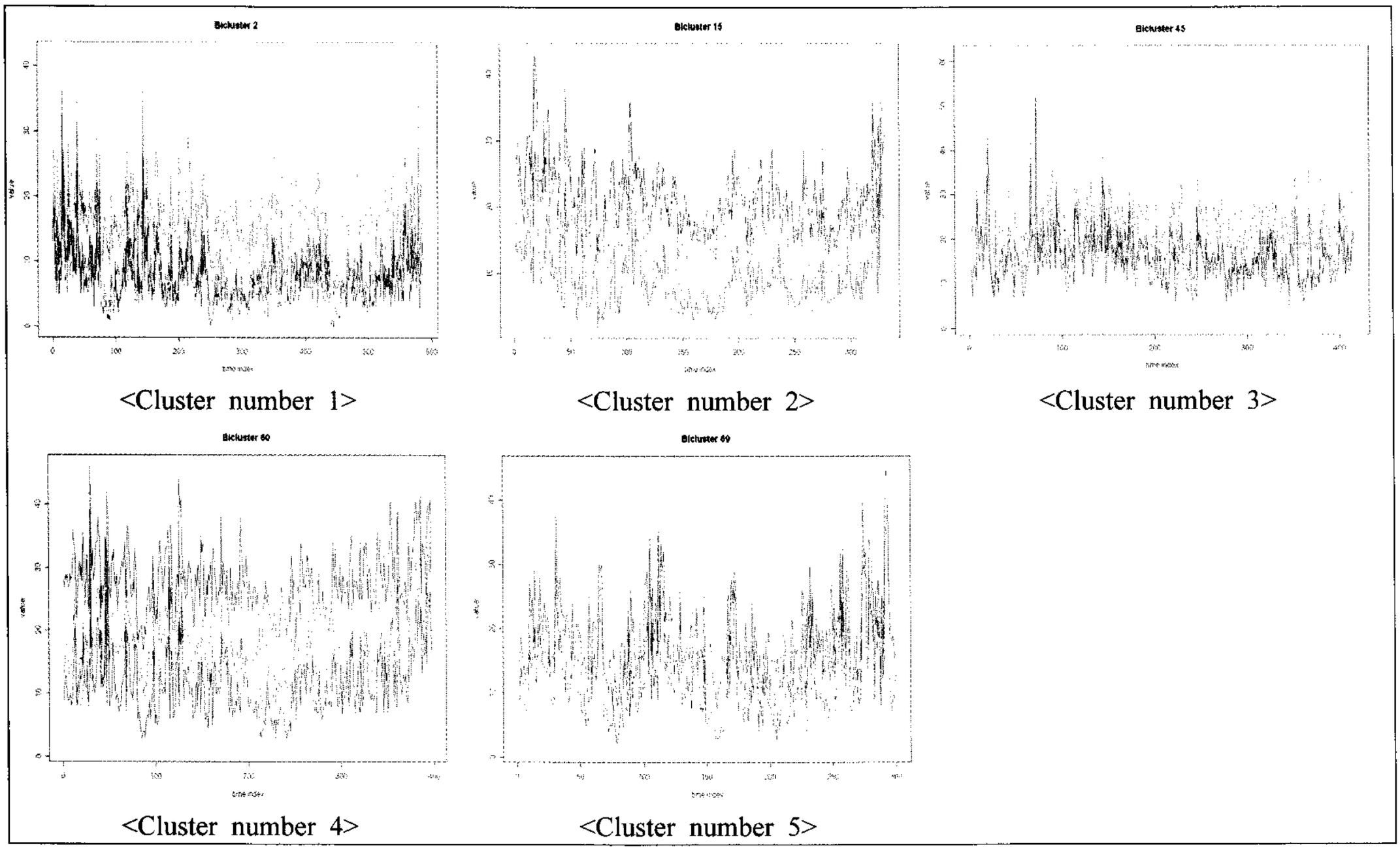
<Cluster number 4> <Cluster number 5>

Fig. 2. Variation of cluster area and cluster time of SO_2.

Table 1. Result of cluster area and cluster time of SO_2

Cluster number	Cluster area	Number of cluster area	Number of cluster time	Frequency (%) of cluster time	H-value
1	Ulsan Deokshinri Chungju Mumhwadong	2	844	9.63	0.0000
2	Ulsan Hyomundong Gangneung Okcheondong	2	877	10.01	0.0137
3	Yeosu Kwangmudong Pohang Jangheongdong	2	5164	58.95	3.8683
4	Chungju Mumhwadong Siheong Sihwa Complex	2	5284	60.32	2.2165
5	Ulsan Sinjeongdong Gangneung Okcheondong	2	3310	37.79	0.2788

울산시 상남리, 포항시 죽도동, 포항시 대도동, 포항시 대송면, 충주시 문화동에서 6.69%의 유사성을 보였으며, 2집단인 울산시 덕신리, 울산시 상남리, 포항시 죽도동, 포항시 대도동, 충주시 문화동은 3.77%, 3집단은 포항시 죽도동, 포항시 대도동, 포항시 대송면, 강릉시 옥천동에서 4.74%, 4집단인 포항시 장흥동, 포항시 죽도동, 포항시 대송면, 충주시 문화동에서 4.53%, 5집단인 여수시 삼일동, 포항시 죽도



<Cluster number 1> <Cluster number 2> <Cluster number 3>

<Cluster number 4> <Cluster number 5>

Fig. 3. Variation of cluster area and cluster time of NO_2.

Table 2. Result of Cluster Area and Cluster time of NO_2

Cluster number	Cluster area	Number of cluster area	Number of cluster time	Frequency (%) of cluster time	H-value
1	Ulsan Sangnamri Pohang Gukdodong Daedodong Pohang Daesongmeon Chungju Mumhwadong	5	586	6.69	2.9665
2	Ulsan Deokshinri Sangnamri Pohang Gukdodong Daedodong Chungju Mumhwadong	5	330	3.77	2.7399
3	Pohang Gukdodong Daedodong Pohang Daesongmun Gangneung Okcheondong	4	415	4.74	2.4563
4	Pohang Jangheongdong Gukdodong Pohang Daesongmun Chungju Mumhwadong	4	397	4.53	2.5373
5	Yeosu Samildong Pohang Gukdodong Daesongmun Gangneung Okcheondong	4	300	3.42	2.5735

동, 포항시 대송면, 강릉시 옥천동에서 3.42% 유사성이 보였다. 다른 물질에 비해 군집 시간대가 낮은 것은 지역간의 이질적 변동성이 강해 보이는 것으로 분석되었는데, 이는 지역간에 유사성이 거의 없

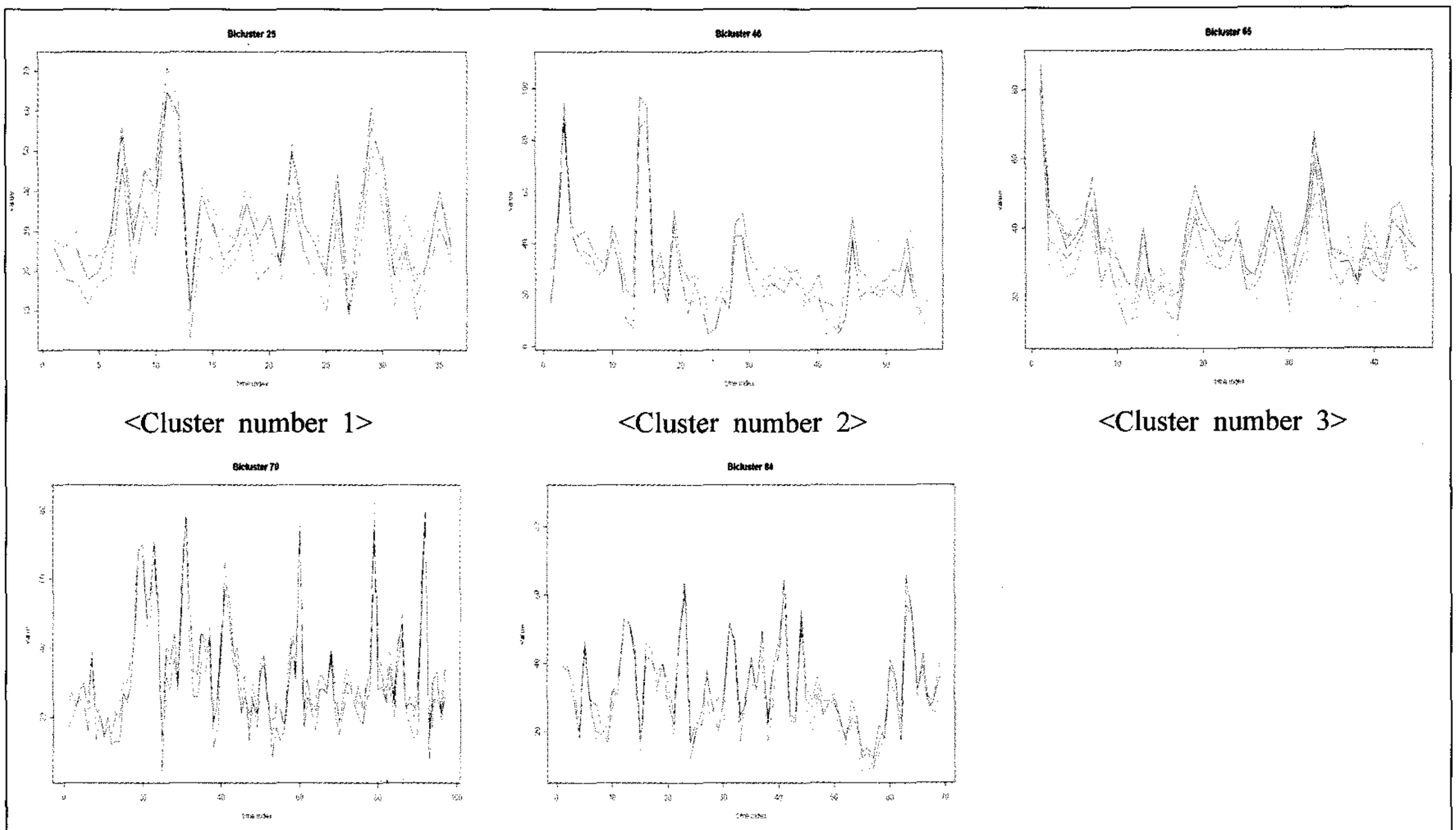


Fig. 4. Variation of cluster area and cluster time of PM_{10}.

Table 3. Result of cluster area and cluster time of PM_{10}

Cluster number	Cluster area	Number of cluster area	Number of cluster time	Frequency (%) of cluster time	H-value
1	Ulsan Hyomundong NongSo 1 dong Yeosu Samildong	3	36	0.41	4.0057
2	Ulsan Hyomundong Daesongdong Yeosu Samildong Pohang Daesongmeon	4	56	0.63	3.7740
3	Ulsan Hwasanri Daesongdong Yeosu Samildong Pohang Daedodong Gangneung Okcheondong	5	45	0.51	4.5759
4	Yeosu Samildong Pohang Daesongmeon Gangneung Okcheondong Ulsan Daesongdong	4	97	1.11	3.7784
5	Ulsan Hwasanri Yeosu Samildong Pohang Jangheongdong Gangneung Okcheondong	4	69	0.78	3.5433

음을 의미한다.

3.3. PM_{10} 물질 분석 결과

대기중 PM_{10} 농도의 분석결과를 Fig. 4와 Table 3에 정리하였다. 1집단으로 울산시 효문동, 울산시 농소1동, 여수시 삼일동, 2집단으로 울산시 효문동, 여수시 삼일동, 포항시 대송면, 울산시 대송동, 3집단은 울산시 화산리, 여수시 삼일동, 포항시 대도동, 강릉시 옥천동, 울산시 대송동 4집단은 여수시 삼일동, 포항시 대송면, 강릉시 옥천동, 울산시 대송동, 5집단은 울산시 화산리, 여수시 삼일동, 포항시 장흥동, 강릉시 옥천동이다. 대기중 PM_{10} 농도에 대한 군집결과, 대체로 매우 낮은 군집시간을 보였으며 각 지역에 따라 PM_{10}농도는 매우 이질적으로 영향을 미치고 있음을 알 수 있다. 즉 대상 지역간에는 유사성이 거의 없었다.

4. 고 찰

일반적으로 대기오염물질의 변동성을 보기위해 가로축을 시간으로 놓고 세로축을 농도로 하여 지역들 간에 시각적으로 평가한 연구는 흔히 볼 수 있다. 다만 이러한 시각적인 자료만을 이용하여 지역들간에 변동성의 차이를 설명하는 것은 한계가 있다. 이러한 고민에 기초하여 어떻게 하면 통계적 분석방법을 통해 수치적으로 나타낼 수 있을까 하는 점에서 본 연구를 접근하였으며 본 연구에서 활용한 통계기법을 활용한 선행연구가 전무한 실정이다.

본 연구에서 지역간 대기환경특성을 비교하기 위해 대기오염 자동측정자료를 이용하였는데, 산업단지 지역의 측정소는 도시지역 전반에 고루 분포한 반면(시화반월 10개소, 울산 13개소, 포항 4개소, 광양만(여수) 3개소), 충주에는 2006년까지 문화동에 1개소만이 운영되고 있어 충주의 대기환경을 대표하는 데는 한계가 있다. 다만 이들 측정자료는 지역별 대기오염현황을 평가할 수 있는 가장 객관적인 자료라는 점에서 기초자료로 활용하였고 지역간 전반적인 대기환경 현황을 비교하는데 이용할 수 있을 것으로 판단된다.

대기오염물질에 직접적인 영향을 미치는 요인은 산업활동, 난방, 자동차운행, 도시개발 등을 들수 있으며, 특히 PM_{10}과 같은 미세먼지농도는 산업체, 자동차 등 배출원에 의한 영향이외에 봄철 황사에 직접적인 영향을 받고 있는데, 기존 연구[11]에 따르면 황사발생일수와 농도가 지역간에 큰 차이가 있음이 지적되고 있다.

본 연구에 이은 향후 과제로 지역간에 유사성을 판단할 수 있는 기준빈도율(Frequency (%) of Cluster Time)을 구하는 것이 필요하다. SO_2는 NO_2와 PM_{10}에 비해 높은 빈도율을 보여주었기 때문에 그리 문제가 되어 보이지 않지만, 현재까지 선행연구가 전무한 상태에서 본 연구결과에 기초하여 바로 기준빈도율을 설정하기란 무리라 여겨진다. 이와 같은 유사성을 판단할 수 있는 유사빈도기준율은 향후 많은 관련 연구를 통해서 경험적 연구에 의해 나올 수 있을 것으로 사료된다.

5. 결 론

본 연구에서는 SO_2, NO_2, PM_{10} 3개 항목의 대기오염물질을 대상으로 환경오염에 따른 건강영향에 관한 조사가 이루어지고 있는 산업단지지역과 대조지역의 지역간 유사한 변동을 가지고 있는가에 대해 통계적으로 분석하였고 다음과 같은 결론을 얻었다.

1) SO_2 농도는 산업단지지역과 대조지역간에 매우 유사한 변동성을 가지는 것으로 분석되었다. 이는 과거 대기중 SO_2 환경기준을 초과하는 경우가 빈번하였고, 지역간 산업활동과 사용연료의 차이가 상당하였으나, 최근에는 전국 대부분의 도시들이 환경기준을 만족하고 있으며 과거 SO_2의 주배출원인이었던 황(S)성분이 높은 석탄, 석유등 화석연료를 대신하여 저유황유와 LNG와 같은 청정연료의 사용이 전국적으로 비교적 고르게 확대됨에 따라 지역간의 대기중 SO_2의 변동성이 유사하게 나타나게 된 원인으로 추정된다.

2) NO_2 농도는 산업단지지역과 대조지역간에 매우 이질적 변동을 가지고 있는 것으로 분석되었다. 즉 지역간에 유사성이 거의 없다고 보이는데, 그 이유는 NO_2의 주요 배출원이 각종 산업활동 및 난방에 따른 연소가스와 자동차 배출가스에 의한 것으로 각 지역의 산업활동 규모와 특성, 자동차운행량

의 차이 등에 따른 것으로 추측된다.

3) PM_{10} 농도 역시 산업단지지역과 대조지역간에 매우 이질적인 변동을 가지고 있는 것으로 분석되었다. 즉 이들 지역간에 유사성이 거의 없다는 결과를 얻었는데, 이는 대기중 PM_{10} 농도에 직접적인 영향을 미치는 황사발생일과 농도의 지역간 편차가 크기 때문으로 여겨진다.

4) 본 연구를 통해 바이클러스터 알고리즘을 이용하여 통계적으로 분석해 본 결과 SO_2는 지역간의 매우 동질적인 변동을 보인 반면 NO_2와 PM_{10}은 매우 이질적인 변동을 보여주고 있다.

참 고 문 헌

1) 지역주민 환경오염노출수준 및 건강영향모니터링(광양, 울산, 포항, 시화 · 반월, 강릉), 2003-2006, 국립환경과학원.
2) 정우식, 이화운, 임헌호, 2007, 바람권역 구분을 통한 부산지역 국지바람 분석, 한국환경과학회지, 16(1), 103-119.
3) 대기환경자동측정자료, 2004-2006, 환경부.
4) 신동천, 2007, 미세먼지의 건강영향, 대한의사협회지, 175-181pp.
5) Hartog J. J., Hoek G., Peters A., Timonen K. L., lbald-Mulli A., Brunekreef B., Heinrich J. P., Van Wijnen J. H., Kreyling W., Kulmala M., Pekkanen J., 2003, Effects of fine and ultrafine particles on cardiorespiratory symptoms in elderly subjects with coronary heart disease, American Journal of Epidemiology, 613-623.
6) Cheng Y., Church G. M., 2000, Biclustering of expression data, In Proc. ISMB'00, AAAI Press, 93-103pp.
7) 성웅현, 1997, 응용다변량분석, 탐진출판사, 321-347pp.
8) Kluger Y., Barsi R., Cheng J. T., Gerstein M., 2003, Spectral biclustering of microarray data: coclustering genes and conditions, Genome Research, 13(4), 703-716.
9) Sara C., Arlindo L., 2004, Biclustering Algorithms for Biological Data Analysis, IEEE Transactions on Computational Biology and Bioinformatics, 1(1), 24-45.
10) 정세균, 김수진, 장병탁, 2006, 유전자 발현 분석을 위한 공진화적 바이클러스터링 기법, 한국컴퓨터종합학술대회 논문집, 33, 1(A), 22-24.
11) 김선영, 이승호, 2006, 한국의 황사발생빈도 분포와 변화분석, 환경영향평가, 15(3), 207-215.