韓國大氣保全學會誌 第 6 卷 第 1 號
J. KAPRA Vol. 6, No. 1 (1990) pp. 1~10

Pattern Recognition을 이용한 지하상가에서의 대기오염물질의 농도 분석에 관한 연구

The Air Quality Analysis in Underground Shopping Centers Using Pattern Recognition

김동술 · 김형석[1)]

경희대학교 자연과학대학 환경학과
[1)] 경희대학교 의과대학 예방의학교실
(원고접수 : 1990. 1. 9)

Dong-Sool Kim, Hyung-Suk Kim[1)]

Dept. of Environmental Science, Kyung Hee University
[1)] Dept. of Preventive Medicine, Kyung Hee University
(Received 9 January 1990)

Abstract

The purpose of the study was to analyze air quality in underground shopping centers using pattern recognition methods. In order to perform this, the concentraion of air pollutants such as CO, NO_2, NO_x, SO_2, and particulate matters was measured at the 11 different shopping centers in Seoul metropolitan area and the total of 47 samples were obtained at random based on the size of shopping centers. To introduce a new concept of the "average concentration" for the indoor air quality analyses, the various multivariate statistical analyses have been studied. Thus, a cluster analysis was applied to separate the samples into pseudo−patterns and a disjoint principal component analysis was used to generate homogeneous patterns after removing outliers from the pseudo−patterns. The 6 homogeneous patterns were then obtained as follows: the first pattern was a group of clean sites; the second a group of sites having high dust concentration; the third a group of sites having high dust and NO_x concentration; the fourth a group of sites having low dust and SO_2 concentraion and high CO concentration; the fifth a group of sites having high NO_2 and SO_2 concentration; and the final a group of miscellaneous sites. Thus, the average concentration could be estimated for each pattern.

1. 서 론

도시의 인구과밀, 지가상승, 상권 확대및 교통난 등의 현상은 시민들의 도심지 지하생활에의 기회를 증대시켰다. 또한 겨울철 에너지 절약의 한 이유로서, 실온유지를 위한 자연적, 인공적 환풍의 부족은 지하공간을 이용하고 있는 다수인에게 실내 대기오염에 대한 관심도를 높여 주었다. 하지만, 우리나라의 경우, 최근까지 실내오염의 기준은 마련되어있지 못하며, 기준설정을 위한 기초 연구 역시 미진한 형편이다.

실내외 오염물의 농도 측정에는 많은 어려움이 따른다. 지하상가의 경우 건축자재, 면적, 문의 개폐상태, 유동인구, 지상오염정도, 시료의 채취장소 및 높이, 건축물의 보수상태 및 연령, 난방 및 조리를 위한 열원의 유무, 열원의 종류, 측정 계절 및 시간, 실외 기상상태, 환풍의 정도, 공기정화기의 유무 및 가동여부에 따라, 실내오염농도는 큰 차이를 보일 수 있다. 또한, 같은 지점, 같은 조건하에서, 시료를 채취 및 분석하더라도, 사용된 기자재의 종류, 실험방법에 따라, 오염물의 농도는 큰 차이를 보일 수 있다. 이와 같은 제약성 때문에, 많은 환경학자들은 가능한 많은 횟수의 시료 채취를 시행하며, 분석방법으로는 "평균"이라는 개념을 많이 이용하고 있다. 하지만, 이러한 개념을 이용할 때에는 주의가 필요하며 통계적 오류를 범하지 않아야 한다.

본 연구는 실내오염물의 농도를 새로운 각도에서 분석하기 위하여, 또한 합리적인 실내대기오염의 통제를 위하여, 서울시내의 지하상가 및 지하통로에서 CO_2, NO_2, NO_x, SO_2 및 분진(dust)을 채집하여 농도 분석을 하였다. 또한, 여기서 양산된 자료는 각종 응용통계를 이용하여 pattern 분류를 시도하였으며, 분류된 각 pattern은 확률적으로 검증한 후, pattern 별로 평균오염농도를 산출하였다.

2. 실험 방법

본 연구는 1988년 1월 6일부터 14일까지, 서울시내 11개소의 지하상가 및 지하통로에서 4가지 gas 상태의 오염물질(CO_2, NO_2, NO_x, SO_2) 및 분진(dust)을 채집하여 농도를 측정하였다. 상가의 규모에 따라, 업소분포에 따라 한 지하상가에서 여러번의 시료채취를 시행하여, 총 47개의 독립된 자료를 얻었다. 또한 각 시료 채취장소의 특이 사항, 즉, 문의 개폐여부, 열원의 종류, 유동 및 상주인구, 환풍 또는 공기정화기의 유무 등을 조사하였다. 각 지점에서 시료 채취는 지상 1m 높이에서 시행하였으며, 시료의 이화학 분석법은 다음과 같았다.

2.1 SO_2 측정

Pararosaniline Formalin법으로 측정하였다. 즉 흡수관에 흡수액 10㎖를 넣은 후 1.5 1/min의 속도로 공기를 통과시켜 SO_2를 흡수시킨 후, pararosaniline formalin 용액 2㎖를 넣고 발색시킨 후 spectrophotometer로 파장 550nm에서 흡광도를 측정하였다.

2.2 NO_2 및 총질소산화물(NO_x)

흡수관에 흡수액(0.01N NH_4OH) 20㎖를 넣고, 유속 1.5 1/min의 속도로 통과시켜 질소산화물을 흡수시킨 후 sulfanilamide 용액 1㎖, naphthyl ethylenediamine 용액 1㎖를 가하여 섞은 후 30분 동안 방치한 다음 spectrophotometer로 파장 535nm에서 흡광도를 측정하였다.

2.3 CO 측정

미국 Ecolyzer 제품인 CO monitor 200을 사용하였다.

2.4 분 진

일본 Kanomax사 제품인 분진계 Model 5300을 사용하였다.

3. 응용통계의 이용 및 결과

3.1 Data의 구조

1988년 1월 중, 서울 시내 11개소의 지하상가 및 지하통로에서 CO, NO_2, NO_x, SO_2, 및 분진을 채집 분석하여 47개의 원자료(raw data)를 얻었다. 도표1은 무작위로 선정된 각 지점별 오염농도와 온도를 보여주고 있다. 오염농도는 같은 지하상가라 할지라도, 농도의 큰 기복을 보였으며, 온도의 경우에는, 외부의 낮은 기온과 관계없는 9−23℃의 온도를 보였다. 이화학 분석된 47개의 시료는 분진의 경우 0.02−0.25mg/m^3, CO의 경우 1−28ppm, NO_2의 경우 0.027−0.104ppm, NO_x의 경우 0.040−0.241ppm, SO_2의 경우 0.010−0.026ppm 범위의 농도를 보여 주었다.

3.2 군집분석법(Cluster Analysis)

군집분석법은 자연과학 분야에서 널리 사용

되고 있는 응용통계 분석법으로서 data 집단에서 유사한 성질을 갖는 특정 group을 분류하는데 이용되고 있다. 환경과학분야에서도 이러한 군집분석법이 보편적으로 이용되어지고 있으며, Hopke[1] 등은 보스턴 지역의 도시분진을 분류하는데 19개의 원소 농도를 변수로하여 군집분석법을 응용한 바 있다. 군집분석법은 사전에 어떤 정보없이 혹은 최소의 알고 있는 정보로 유용한 집단을 분류 및 확인할 수 있으며, 대규모 data에서 새로운 개념을 갖게하여 준다. 이 분석법의 기본 원리는 두 objects 사이의 거리를 기준으로 공간에서 비유사도(dissimilarity)를 측정하므로서, 서로 인접해 있는 object를 모아 들이는데 있다. 일반적으로 군집분석법에는 크게 두가지 방법이 있다. 즉, 위계(hierarchical)분석법과 비위계(non-hierarchical)분석법으로 나눌 수 있다. 위계분석법은 응집위계(agglomerative hierarchical)와 분산위계(divisive hierarchical)분석법으로 세분된다. 응집위계법은 각 object에서 군집(cluster)을 만들기 시작해서 조그만 군집들이 마치 나무의 가지를 치듯이 하나의 큰 군집을 만들면서 끝을 맺는다. 반면에, 분산위계법은 응집위계법의 역과정으로 생각할 수 있다. 즉, 하나의 큰 군집이 세분되기 시작하여 한 개의 object가 한 개의 군집을 만들때까지 분열되는 분석법이다. 일반적으로, 비위계분석법은 특정 군집의 숫자를 알거나, 가정할 수 있을 때, 전체 object들을 주어진 특정 군집수로 최적 분배하는 방법이다. 위계분석법의 장점은 비위계분석법에 비해 algorithm이 간단하고, computer 시간이 상대적으로 짧은데 있다. 하지만, 군집분석법은 연구과정에서 자료의 형태 윤곽은 쉽게 파악할 수 있으나, 연구의 최종 결과를 얻고자 할 때 많은 주의가 필요하다. 위계 및 비위계 군집분석법에는 수많은 algorithm이 있으며 많은 문헌에서 이들의 특성을 찾아 볼 수 있다[2-6].

AGCLUS

AGCLUS는 FORTRAN IV로 쓰여진 Computer Program[7]으로, 응집위계군집분석법(agglomerative hierarchical cluster analysis)을 주 사용 목적으로 하며, 비유사도 측정을 위해 7가지 사양(option)을 연구자에게 제공하고 있다. 비유사도의 선택은 군집의 크기와 군집 간의 거리를 서로 다르게 나타낼 수 있으므로, 비유사도 선택을 위해 응용에 앞선 충분한 사전 연구가 필요하다. 7가지 사양에 대한 수학적인 표현과 특성은 Hopke에 의해 서술된 바 있다[8].

AGCLUS의 응용 결과

AGCLUS를 이용하여 서울시 지하상가에서 얻은 47개의 자료를 분류하였다. 이러한 군집분석을 할 경우, 원자료가 지니고 있는 물리화학적 변수의 선정(variable selection)과 자료의 변환(data transformation)이 중요하다. 왜냐하면, 특정변수의 변량이 농도에 대해 분포도에서 치우쳐(skewed)있을 경우 및 변수사이에 심한 상관관계를 보일 경우, 군집분석의 결과는 과장되거나, 오류를 범할 수 있기 때문이다. 따라서, 변량의 분산분포조사와 변수사이의 상관관계조사는 군집분석에 앞서 선행되어야 한다. 변량의 분포정도가 치우쳐 있거나, 상관관계가 발견되었을 경우, 분산분포는 최소한의 대칭형이 되도록 자료변환을 해야하며, 심한 상관관계를 보이는 변수는 분류작업에서 제외시켜야 한다.

본 연구를 위하여, 처음에는 온도변수를 포함한, 5개의 농도변수(dust, CO, NO_2, NO_X, SO_2)를 전부 분류작업에 이용하였다. 이 경우 수상도(dendrogram)는 복잡한 양상을 보여 분류를 효율적으로 수행할 수 없었다. 그러므로, 온도변수를 제외하고 5개의 농도변수만을 이용하여 재분류하였다. 분류작업에 앞서, 각 변수의 분산 정도를 파악하였으며, 변수사이의 상관관계를 살펴보았다. 각 변수에 내한 변량의 분산정도는 양과 음의 방향 어느곳으로도 치우쳐 있지 않았으며, 변수간의 상관관계도 관찰되지 않았다. 특히 우려했던 NO_2와 NO_X 사이의 상관관계도 독립됨을 보였다. 따라서, 분류분석에 앞선, 자료의 변환은 시도되지 않았다.

AGCLUS를 이용한 원자료의 분류를 위해, AGCLUS속의 사양을 이용하여, 원자료를 표준화(standardization) 시킨 후, 5가지의 서로 다른 비유사도(dissimilarity)에 의해, 수상도를

작성하였다. 특정 응용에 맞는 올바른 비유사도의 선택이 위계분석법 응용에서는 중요하다. Hopke등[9]은 환경 관련 자료로부터 해석 가능한 군집 분류를 하기 위해 Euclidian 거리의 제곱평균을 비유사도로 사용했다. 즉, 군집 속에서 object간 거리의 제곱의 합이 최소 증가를 보일 때, 이미 군집화된 자료를 해석하는데 가장 유용함을 보인 바 있다. 하지만, 본 연구의 경우, Euclidian 거리의 제곱평균보다 평균차이(MD : mean difference)를 비유사도로 사용할 때, 지하상가의 자료를 쉽게 해석할 수 있었다. 평균차이(MD)는 수학적으로 다음과 같이 표시할 수 있다. 즉, object j와 k 사이의 평균차이는 변수 i에 대하여,

$$MD_{jk}=\left[\sum_{i=1}^{m}(x_{ji}-x_{ki})\right]/m$$

그림 1은 평균차이를 비유사도로 이용한 수상도(dendrogram)이다. 그림에서 보는 바와 같이, 수상도의 처음 윗부분은 많은 object들이 비교적 낮은 비유사도 준위에서 질서있게 모여 있는 것을 볼 수 있으며, 밑으로 갈수록 적은 object들이 높은 비유사도 준위(level)에서, 산만하게 모여 있는 것을 볼 수 있다. 만약, 이 수상도에서 이론적 기준에 의하여 최적 비유사도 준위를 결정할 수 있다면, 지하상가에서의 오염 pattern은 쉽게 파악할 수 있다. 하지만, 최근까지 위계군집분석법의 수상도에서 최적 비유사도 준위의 결정방법과, 비위계군집분석법에서 최적 군집수의 결정방법은 통계학적 견지에서 상당히 주관적이다. 따라서, 군집분석법을 이용하여 유용한 정보는 얻을 수 있으나, 연구 주제에 관한 양적 결과를 얻기에는 어려운 면이 있다. 이러한 문제점에 대한 해결책으로 주인자분석법(principal component analysis), 요인분석법(factor analysis) 등과 같은 다변수통계학(multivariate statistics)을 동원하여, 이미 선택된 군집 내의 object들을 확률적으로 검증할 수 있다.

따라서, 본 연구는 그림 1의 수상도를 이용하여, 비교적 낮은 비유사도 준위를 주관적으로 설정한 후, 잠정적 pattern 분류를 시도하였다. 즉, 주어진 비유사도에서, class 1은 14개의

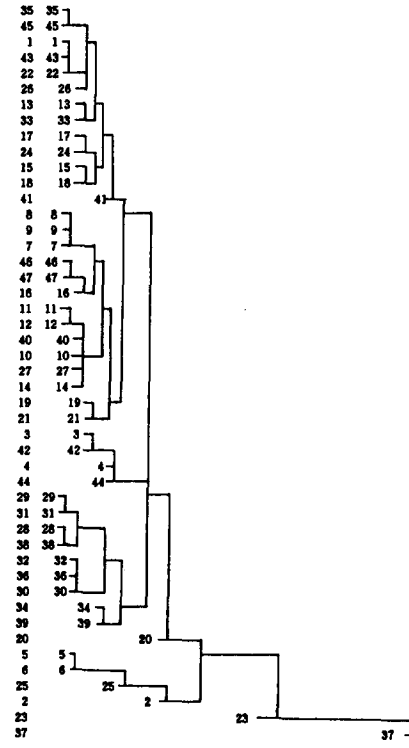

Fig. 1. A dendrogram for samples obtained from underground shopping centers in Seoul.

objects(8, 9, 7, 46, 47, 16, 11, 12, 40, 10, 27, 14, 19, 21), class 2개는 8개의 objects(35, 45, 1, 43, 22, 26, 13, 33), class 3은 4개의 objects(17, 24, 15, 18), class 4는 4개의 objects (3, 42, 4, 44), class 5는 9개의 objects(29, 31, 28, 38, 32, 36, 30, 34, 39) 및 기타 class 99의 8개 objects 등으로 잠정적인 가분류를 시행할 수 있었다.

3.3 주인자분석법(Principal Component Analysis)

군집분석법을 이용하여 유사한 object들이 각각의 군집을 형성하고 미완성이지만 일단 object들의 분류작업이 끝났을 때, 각 군집에 분류된 object들이 과연 그 class(또는 pattern)에 속해 있는지 혹은 아닌지를 양적으로 확인

할 수 있는 방법이 필요하다. 이와 같은 연구를 수행하기 위해, 분리주인자분석법(disjoint principal component analysis)을 응용하였으며, 이를 위해 SIMCA Package[10]를 사용하였다. 주인자분석법의 기본개념은 다음과 같다. 공간에서 한개에서 세개까지의 변수가 존재한다면, 각 object 사이의 관계는 한개의 공간좌표 그림으로 쉽게 이해할 수 있다. 만약 세개 이상의 변수가 있다면, object 사이의 관계를 시각적으로 이해하기는 불가능하다. 하지만 변수의 차원을 줄일 수만 있다면 이들의 관계를 쉽게 이해할 수 있다. 이와 같은 차원축소법은 정사영(orthogonal projection)과 새로운 변수로 재구성된 축소 공간의 축에 의해 수행되며, 새로운 변수들은 본래 변수들과 선형조화(linear combination)된다. 이 분석과정을 주인자분석법, 또는 eigenvector analysis라고 한다. 주인자분석에서 중요한 점은 통계학적으로 의미있는 주인자(principal components)의 숫자를 결정하여, 차원 축소를 주인자 숫자에 준해 수행하는 것이다.

SIMCA(Soft Independent Modeling of Class Analogy)

SIMCA는 Wold[11]에 의해 개발되었으며, SIMCA-3B[10]는 microcomputer를 위한 Basic언어로 표시되어 있다. SIMCA 분석에 앞서, 몇가지 용어 설명이 필요하다. 각 class들이 미리 알고 있는 정보에 의해, 혹은 군집분석과 같은 방법으로 임의 분류될 수 있을 때, 이 class를 training set라고 하며, 어떤 class로도 분류되지는 않았지만 분류되어야 할, 소속이 불확실한 object들의 모임을 test set라고 한다. 또한 modeling 후 어떤 class에도 속하지 않는 object를 특이점(outlier)라고 한다. SIMCA에 의한 자료분석은 두단계로 나눌 수 있다. 첫단계는 training set를 이용하여 주인자 model을 개발하는 것이고, 두번째 단계는 test set중의 각 object를 이미 개발된 각 class model과 비교해서 소속감(membership)을 부여하는 것이다. 이 SIMCA를 응용할 경우 군집분석에서와 마찬가지로 자료의 변환은 중요하다. 만약 object 중의 한 변수가 다른 변수에 비해 치우쳐 있다면, 적당한 변환을 수행하여 그 변수의 과잉영향을 줄일 수 있다. 일반적으로, 주인자분석을 할 경우, 인자(component)의 숫자가 증가할수록, 그 모델은 더 좋은 fitting을 하지만, 결과에 대한 유효성(validity)은 더 큰 제약을 받게 된다. 따라서 모델에 대한 유효성 검증이 중요하며, 모델의 최적화(optimization)를 위해 횡유효도(cross validation) 검사를 검증의 도구로 이용하고 있다[12]. SIMCA를 이용한 분류원리는 다음과 같다. 일단, 각 class가 선형구조로 모델화되고, 한개의 object가 고정된 확률값에서 이미 모델화 된 특정 class속에 소속할 수 있는지를 임계거리(critical distance)를 이용하여 결정하는 것이다. SIMCA의 원리는 다중 Taylor 급수(multiple Taylor's expansion)에서부터 출발하는데, 이는 몇 개의 분산계수(discrete parameter)인 평균치 y, 변수 관련 항 b, object의 값 t로 표시된다.

$$y_{ij}^{(q)} = y_i^{(q)} + \sum_k b_{ik}^{(q)}\, t_{kj}^{(q)} + e_{ij}^{(q)} \quad \cdots\cdots (1)$$

Data 횡열 y_{ij}는 q번째 class에서 k개의 주인자를 가진 i번째 변수와 j번째 object의 측정치를 의미한다. 각 계수 y, b, t는 분류계산을 위해 한 file 속에 저장된다. 수식 (1)에서, e_{ij}의 제곱의 합은 class model과 그 class에 속해있는 한 object 사이의 거리를 나타낸다. 따라서 class q의 잔여표준편차(residual standard deviation)는 다음과 같이 주어질 수 있다.

$$S_o^{(q)} = \left[\sum_i \sum_j e_{ij}^{2(q)} / (n_q - a_q - 1)(m - a_q) \right]^{1/2} \quad \cdots\cdots (2)$$

여기서 n_q, a_q, m은 각각 class q내의 object 수, 주인자의 수 및 변수의 수를 의미한다. Test set에서 소속이 불확실한 object p의 분류를 위하여, 각 class q의 model을 이용하며, 일반 다중회귀분석(multiple linear regression analysis)이 적용된다.

$$y_{ip} - y_i^{(q)} = \sum_k b_{ik}^{(q)}\, t_{kp}^{(q)} + e_{ip}^{(q)} \quad \cdots\cdots (3)$$

잔여 e_{ip}는 object p의 분류기준으로 사용될 수 있다. 따라서 class q에 대한 object p의 잔

여표준편차는 다음과 같다.

$$s_p^{(q)}=[\sum_i e_{ip}^{2(q)}/(m-a_q)]^{1/2} \quad \text{.....(2)}$$

만약 잔여표준편차 $s_p^{(q)}$가 class q의 잔여표준편차 $s_o^{(q)}$보다 작으면, object p는 class q로 분류된다. 이와같이 소속이 확실하지 않은 object는 모델화된 각 class와 비교되어 가장 가까운 class에 속하게 되는데, 이는 F-test에 의해 수행된다.

$$F=\Phi^2 s_p^{2(q)}/s_o^{2(q)} \quad \text{.....(5)}$$

여기서 Φ는 class q의 주인자 model에 대한 보정계수(correction factor)가 된다.

$$\Phi^2=[n_q/(n_q-a_q-1)] \quad \text{.....(6)}$$

주어진 level에서 F의 임계값은 $(m-a_q)$와 $(n_q-a_q-1)(m-a_q)$의 자유도(degree of freedom)를 갖는다. 한개의 training set가 두개 이상의 class로 구성되어 있을 때, 각 class는 표준결정도(standard decision plot: Cooman, et al.[13])를 이용하여 시각적으로 비교할 수 있다.

SIMCA의 응용결과

SIMCA는 많은 분야에서 응용되고 있다. Kvalheim[14]은 환경연구를 위해 capillary gas chromatography를 이용하여 50−60개의 주 peak를 자료로하여 10개의 bluemussel의 조직분류를 하였고, Scott[15]는 가상 대기오염 자료를 가지고 자료의 윤곽 파악 및 특이점 검출에 SIMCA를 이용하였다.

앞에서, 위계군집분석법의 수상도를 이용하여, 잠정적으로 기타 class를 포함한 6개의 class를 만들었다. 본 연구의 경우, 소속감을 부여할 test sets가 존재하지 않으므로, SIMCA의 첫번째 응용단계로서 training sets를 이용한 주인자 모델을 개발하였다. 우선, 각 class 속의 object가 과연 class 속에 확률적으로 무리없이 존재할 수 있는지를 검사하였다. 이를 위해, 이미 잠정적으로 분류된, class 1의 14개, class 2의 8개, class 3의 4개, class 4의 4개, class 5의 9개 및 기타 class 99의 8개 objects를 각각 training set로 하고, 분리주인자분석을 시행하였으며, 총 6개의 sets를 모델링한 결과, 어느 class도 단 한개의 주인자를 갖지 않음을 관찰하였다. 식 (2)와 (4)에 의해, 잔여 표준편차를 구할 수 있었으며, 계산된 값을 이용하여, Cooman의 표준결정도(standard decision plot)를 작성할 수 있었다.

그림 2는 class 2와 3을 위한 표준결정도이다. 그림에서, 평면은 95% 확률의 임계거리(critical distance)선에 의해 4개 영역으로 나누어졌다. 그림 왼쪽 영역은 class 2의 object들만이 존재하는 영역이고, 오른쪽 아래 영역은 class 3의 object들만이 존재하는 영역이다. 왼쪽 아래 영역은 두 class가 공존하는 영역이고, 오른쪽 위 영역은 class 2와 3을 제외한 그밖의 class의 object가 존재할 영역이다. 만약 이 그림 위에 그밖의 class 1, 4, 99의 object를 삽입한다면, class 1, 4, 99의 모든 objects는 오른쪽 위 영역에 자리할 것이다. 따라서, class 2와 class 3의 objects가 그림 왼쪽 아래 및 오른쪽 위 영역에 존재할 경우, 분류는 95% 확률로 잘못되었다고 생각할 수 있다.

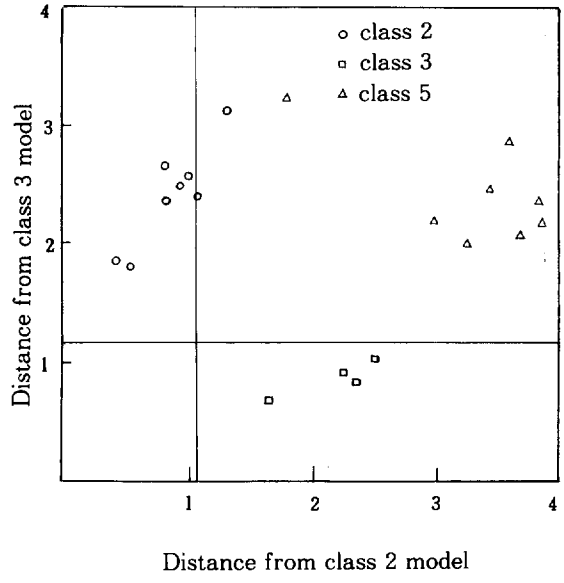

Fig. 2. Decision plot for class 2 and class 3 models.

Table 1. Raw data obtained from various underground shopping places.

ID	Sampling Sites	Temp (℃)	Dust (mg/m^3)	CO (ppm)	NO_2 (ppm)	NO_X (ppm)	SO_2 (ppm)
1	Chungryang Restaurant	20.1	0.14	3	0.0548	0.1002	0.0189
2	Chungryang Coffee	19.7	0.25	7	0.0401	0.0869	0.0125
3	Chungryang Theater	20.7	0.06	13	0.0535	0.0735	0.0200
4	Chungryang Billiard	12.5	0.09	14	0.0301	0.0601	0.0120
5	Jonggak Passage	18.6	0.12	7	0.0354	0.1497	0.0104
6	Jonggak Drugstore	18.7	0.13	3	0.0331	0.1229	0.0119
7	Jonggak Bookstore	19.8	0.07	1	0.0427	0.1003	0.0128
8	Dongdaemoon	10.3	0.06	3	0.0347	0.0962	0.0136
9	Dongdaemoon Shoes	12.2	0.06	4	0.0414	0.1136	0.0136
10	Myungdong Passage	16.4	0.07	6	0.0467	0.0802	0.0180
11	Myungdong Baby Store	19.4	0.03	3	0.0467	0.0802	0.0164
12	Myungdong Records	17.4	0.06	2	0.0434	0.0668	0.0160
13	Kangnam T. Passage	13.6	0.11	2	0.0501	0.0802	0.0214
14	Kangnam T. Passage	21.7	0.11	3	0.0548	0.0869	0.0165
15	Kangnam T. Clothes	20.8	0.13	3	0.0434	0.1336	0.0183
16	Kangnam T. Passage	18.5	0.15	1	0.0354	0.1049	0.0164
17	Kangnam T. Passage	11.1	0.11	5	0.0354	0.1604	0.0216
18	Kangnam T. Restaurant	23.0	0.17	4	0.0501	0.1470	0.0164
19	Kangnam Passage	14.7	0.15	4	0.0274	0.0400	0.0189
20	Kangnam Passage	10.1	0.09	2	0.0327	0.0454	0.0264
21	Kangnam Clothes	19.5	0.08	3	0.0334	0.0467	0.0176
22	Kangnam Restaurant	19.2	0.15	5	0.0601	0.0815	0.0189
23	Kangnam T. Coffee Shop	19.2	0.02	4	0.0614	0.2406	0.0149
24	Kangnam T. Passage	11.3	0.13	2	0.0501	0.1604	0.0214
25	Kangnam T. Passage	21.6	0.19	3	0.0401	0.2005	0.0158
26	Chungryang Chess	14.7	0.19	5	0.0535	0.0909	0.0183
27	Jungro Clothes	11.3	0.08	5	0.0601	0.0869	0.0200
28	Jongo Metal Craftsshop	12.9	0.08	5	0.0802	0.0802	0.0247
29	Jongro Craftsshop	13.2	0.12	8	0.0668	0.0975	0.0252
30	Jongro Pottery Shop	15.6	0.16	8	0.0909	0.1002	0.0228
31	Jongro Yarn Shop	14.5	0.13	10	0.0668	0.0922	0.0246
32	Jongro 5 Credit Union	21.7	0.14	6	0.1036	0.1203	0.0252
33	Jongro 5 Art. Flower	13.3	0.09	2	0.0668	0.0775	0.0228
34	Chungryang Cosmetic	19.9	0.09	3	0.1002	0.1470	0.0211
35	Chungryang Electro Ent.	21.6	0.13	1	0.0735	0.1096	0.0194
36	Chungryang Restraurant	20.1	0.13	3	0.1002	0.1070	0.0217
37	Namdaemoon Art Shop	19.1	0.08	28	0.1002	0.1270	0.0257
38	Namdaemoon Passage	9.1	0.08	6	0.0668	0.0802	0.0237
39	Elgiro Clothes	21.6	0.03	2	0.0802	0.1136	0.0205
40	Elgiro Passage	9.0	0.05	2	0.0635	0.0668	0.0189
41	Elgiro Coffee Shop	18.7	0.23	4	0.0668	0.1270	0.0194
42	Elgiro Food Dept.	23.1	0.05	10	0.0401	0.1042	0.0149
43	Elgiro Clothes	17.5	0.15	4	0.0601	0.1070	0.0189
44	Elgiro Shoes	16.2	0.13	12	0.0735	0.0935	0.0164
45	Elgiro Passage	11.3	0.13	1	0.0701	0.1070	0.0194
46	Pyunghwa 2nd Ground	14.3	0.10	2	0.0334	0.0935	0.0158
47	Pyunghwa 1st Ground	17.2	0.11	5	0.0328	0.1070	0.0149

그림 2에 의하면, 잠정적 class 3 속의 4개의 objects는 완벽하게 class 3으로 분류되었으며, 잠정적 class 2속의 8개 objects 중 하나(Object No. 33)는 기타 class에 속하며, 하나(Object No. 26)는 확률선상에 놓여 있다. 여기서, 하나의 outlier object, 즉 object No. 33을 제거시킨다면, 나머지 7개 objects로 구성된 순수 class 2를 창조할 수 있다. 이와같이 각 평면위에 서로 다른 class를 각각의 임계거리에 준하여 상호비교하여 각 class의 순수 objects와 outlier objects를 분리결정할 수 있다. 즉 각 class의 횡유효도를 이용하여 outlier object를 제거한 후, 잠정적 class를 순수 class로 만들 수 있다.

도표 2는 군집분석에 의해 인위적으로 생성된 5개 class를 SIMCA에 의해 outlier object를 제거한 후 만든 순수 class들이다. 잠정적 class에서 제거된 outlier objects는 상호비교되어 기타 class 99에 포함시켰다. 따라서, 이들 class의 pattern분류는 최종적으로 다음과 같다. 첫째, class 1은 모든 오염물의 농도가 낮은 청정구역으로, 이 구역의 오염농도 평균값은 분진 0.07mg/m^3, CO 2.6ppm, NO_2, 0.039ppm, NO_X

Table 2. A classification result after deleting outliers in each class using a disjoint principal component analysis.

Class	ID	Heating	Door	Average Concentration				
				Dust (mg/m^3)	CO (ppm)	NO_2 (ppm)	NO_X (ppm)	SO_2 (ppm)
1	2	None	None	0.07	2.6	0.039	0.085	0.015
	9	None	Open					
	7	None	Close					
	46	None	None					
	11	None	Close					
	12	None	Open					
	21	None	Close					
2	35	None	Open	0.14	3.0	0.060	0.097	0.019
	45	None	None					
	1	Gas	Close					
	43	Electro	Close					
	22	Gas	Close					
	26	Kerosene	Close					
	13	None	None					
3	17	None	None	0.14	3.5	0.045	0.150	0.019
	24	None	None					
	15	None	Open					
	18	Gas	Open					
4	3	None	Close	0.08	12.3	0.049	0.083	0.016
	42	Gas	Close					
	4	None	Close					
	44	Kerosene	Close					
5	29	Kerosene	Close	0.12	6.6	0.082	0.097	0.024
	31	Kerosene	Open					
	28	Kerosene	Close					
	38	Local	None					
	32	Kerosene	Close					
	36	Gas	Close					
	30	Kerosene	Close					

0.085ppm, SO_2 0.015ppm였다. 이 구역의 경우 난방이나, 취사를 위한 열원은 전혀 없었다. 둘째, class 2는 분진농도가 높은 구역으로 분진의 평균농도가 0.14mg/m^3이었다. 세째, class 3은 분진과 NO_X의 농도가 높은 구역으로 분진과 NO_X의 평균농도가 각각 0.14mg/m^3, 0.15ppm이었다. 네째, class 4는 분진과 SO_2의 농도가 낮으며, CO의 농도가 높은 구역으로 CO의 평균농도가 12.3ppm이었다. 이 구역의 경우 출입문이 모두 닫혀 밀폐되어 있었다. 다섯째, class 5는 NO_2 및 SO_2의 농도가 높은 구역으로 NO_2의 평균농도가 0.082ppm, SO_2의 평균농도가 0.024ppm이었다. 이 구역의 경우, 난방을 위한 연료로서 석유를 대부분 이용하였다. 마지막으로, class 99는 outlier objects를 포함한 기타 구역으로 이 지역에서 산출된 각 오염물의 평균농도는 위의 5개 pattern에서 각각 산출된 평균농도의 범위(range)안에 있었다.

4. 결 론

서울시 지하상가에서 측정된 47개의 자료를 바탕으로, 실내 대기오염물질의 평균농도를 새로운 방법으로 분석 산출하였다. 본 연구를 위해, 군집분석법(cluster analysis)과 분리주인자분석법(disjoint principal component analysis)을 응용하였다. 따라서, 47개의 자료는 기타 class를 포함한 6개의 pattern으로 분류되었고, 각 class 속의 오염물 평균농도는 각 pattern에 포함된 outliers를 95% 확률로 제거한 후 산출하였다. 본 연구에서 얻은 6개 pattern은 다음과 같았다. 첫째, 모든 오염물의 농도가 낮은 청정구역 ; 둘째, 분진의 농도가 높은 구역 ; 세째, 분진과 NO_X의 농도가 높은 구역 ; 분진과 SO_2의 농도가 낮으며 CO의 농도가 높은 구역 ; NO_2 및 SO_2의 농도가 높은 구역 및 여섯째, 기타 구역 등이었다.

참 고 문 헌

1. Hopke P.K., et al, (1976), The Use of Multivariate Analysis to Identify Sources of Selected Elements in the Boston Urban Aerosol, Atmospheric Environment, 10, 1015－1025.
2. Everitt B., (1977), Cluster Analysis, 2nd ed, Halsted Pressn, New York.
3. Hartigan J.A., (1975), Clustering Algorithms, John Wiley & Sons, Inc., New York.
4. Kaufman L., Massart, D.L., (1984), Cluster Analysis-Chemometrics, Kowalski B.R. ed, Nato ASI Series C, Vol. 138, Reidel Publishing Company, Boston.
5. Massart D.L., Kaufman L., (1983), The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, John Wiley & Sons, Inc., New York.
6. Vogt W., Nagel D., Sator H., (1987), Cluster Analysis in Clinical Chemistry, John Wiley & Sons, Inc., Chichester.
7. Oliver D.C., (1973), Aggregative Hierarchical Clustering Program Write-Up, National Bureau of Economic Research, Cambridge, Ma.
8. Hopke P.K., (1983), Introduction to Multivariate Analysis of Environmental Data, Natusch D., Hopke P.K., ed, John Wiley & Sons, Inc., New York.
9. Hopke P.K., (1976), Application of Multivariate Analysis to the Interpretation of the Chemical and Physical Analysis of Lake Sediments, J. Env. Sci. Health, All, 367－383.
10. SIMCA－3B Manual, (1984), A Pattern Recogniton Program for CPM and MS-DOS Based Microcomputers, Principal Data Components, 2505 Shepard Blvd, Columbia, MO.
11. Wold S., (1976), Pattern Recognition by Means of Disjoint Principal Components Models, Pattern Recog., 8, 127.
12. Wold S., (1978), Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models, Technometrics, 20, 397－405.

13. Coomans D., et al., (1981), Pilot Study of the Applicability of the SIMCA Pattern Recog-nition Method to Clinical problems, Patroonherkenning Laboratoriumonderzoe-ken, Hoofdstuk 14.
14. Kvalheim O.M., et al., (1983), SIMCA Multivariate Data Analysis of Blue Mussel Component in Environmental Pollution Studies, Anal. Chim Acta, 150, 145－152.
15. Scott D.R., (1985), Environmental Applica-tion of Chemometrics, Breen J.J., Robinson P.E., ed., ACS.