<Back to Index>

TAO, Vol. 15, No. 2, 199-220, June 2004

Pattern Recognition of High O_3 Episodes in Forecasting Daily Maximum Ozone Levels

Jeong-Sook Heo[1], Ki-Hyun Kim[2], Dong-Sool Kim[1,*]

(Manuscript received 22 December 2003, in final form 9 April 2004)

ABSTRACT

In this study, a method was developed to diagnose ozone episodes exceeding environmental criteria (e.g., above 80 ppb) on the basis of a multivariate statistical method and a fuzzy expert system. This method, being capable of characterizing the occurrence patterns of high-level ozone, was employed to forecast daily maximum ozone levels. The hourly data for both air pollutants and meteorological parameters, obtained both at the surface and at high elevation (500 hPa) stations of Seoul City (1989-1996), were analyzed using this method. Through an application of the fuzzy expert system, the data sets were classified into 8 different types for common ozone episodes. In addition, the data sets were divided into patterns of 11 (Station A), 20 (Station B), 8 (Station C), and 10 (Station D) for site-specific ozone episodes. The results of the analysis were successful in demonstrating that the method was sufficiently efficient to classify each class quantitatively with its own patterns of ozone pollution.

(Key words: Ozone, Multivariate statistical method, Fuzzy expert system, Patterns of high-level ozone.)

1. INTRODUCTION

Elevated tropospheric ozone has been a serious air pollution problem over the past several years in Korea due to its adverse impact on human health, crops, and trees (Ghim and Chang

[1] Department of Environmental Science and Engineering, School of Environment and Applied Chemistry, Kyung-Hee, University Yongin-City, South Korea

[2] Department of Earth Environmental Sciences, Sejong University, Seoul, South Korea

[*] *Corresponding author address:*Dong-Sool Kim, Department of Environmental Science and Engineering, School of Environmental and Applied Chemistry, Kyung-Hee, University Yongin-City, Kyunggi-do 449-701, South Korea; E-mail: atmos@khu.ac.kr

2000). The production of tropospheric ozone is known to be regulated by a variety of natural and anthropogenic processes coupled with diverse meteorological conditions (Vukovich 1995; Chan et al. 1998a and b; Wang et al. 2003a). In many cases, control strategies for ozone levels have been established based on simulation approaches: such as photochemical, statistical and neural network modeling (Hanna et al. 1996; Heo and Kim 2004). Hence, the optimal management tactics for ozone-related pollution require accurate forecasts of high-level ozone occurrences (Robeson and Steyn 1990; Prybutok et al. 2000).

In recognition of the significance of ozone chemistry, many researchers have studied the relationship between high-level ozone and relevant environmental parameters (Comrie 1997; Hadjiiski and Hopke 2000). Many attempts have also been directed at the development of models dependent on statistical analysis of (current and previous) meteorological conditions and precursors of ozone (Chen et al. 1998; Gardner and Dorling 2000; Yu and Chang 2000; Ballester et al. 2002; Wang et al. 2003b). To date, multiple-linear regression models have been widely applied to the prediction of ozone concentrations (Robeson and Steyn 1990; Ryan 1995; Hubbard and Cobourn 1998); and recently, new techniques utilizing fuzzy and neural network models have been developed and demonstrated (Comrie 1997; Gardner and Dorling 2000; Peton et al., 2000; Lu et al. 2002; Chaloulakou et al. 2003; Wang et al. 2003b). However, pollution patterns of ozone are not simple enough to be described solely by such factors as the combined effects of various potential precursors and the prevailing meteorological conditions. Explanation of ozone behavior in these previous models was further limited by a number of factors; for example, site-specificity, and nonlinear relationships between different variables, etc. Thus, design schemes for forecast modeling have been highly diverse, and dependent on the selected patterns of ozone pollution. In well-evaluated empirical modeling, it is necessary to identify the dominant patterns of past high-level ozone episodes, and to obtain their site-specific characteristics. For this reason, this study was conducted on the basis of two statistical methods: cluster and disjoint principal components analysis, and a fuzzy expert system to characterize high-level ozone episodes.

The most common approach to selecting structures for numerical data involves the employment of multivariate statistical methods for pattern recognition including cluster and principal components analysis (Startis et al. 1995). Cluster analysis has been recognized as an important analytical tool in simplifying complex information and identifying tentatively similar patterns (in the whole dataset) without priori information (Dubes and Jain 1979; Hopke 1985; Dorling et al. 1992). However, a drawback to its application involves the subjective selection of the optimum number of clusters. Despite numerous misgivings regarding this problem, its application is still favored; as to date there has been no completely satifactory solution to this problem (Vogt 1987; Huan 1992). When there are more than three dimensions (variables), it is difficult to characterize the relationships among different objects. Thus, a reduction technique for dimensions is needed to simplify relationships with the least system bias. Previously, principal component analysis has been applied successfully to identify dominant multivariate relationships in measured data (Wold 1976; Yu and Chang 2000). This method was used to investigate the feasibility of cluster analysis in class assignments. A fuzzy expert system based on fuzzy logic is able to express the models in terms of fuzzy rules for complex phenomena. Thus, it can facilitate the recognition of patterns in complex nonlinear phenomena (Peton 2000). Its versatility has been demonstrated sufficiently in many different areas of pattern recognition (e.g., Zimmermann 1990).

Based on these statistical approaches, Yoo and Kim (1997) were able to explain particulate pollution phenomena in terms of the relationships between various chemical compounds and weather types. Chung et al. (1996) also attempted to estimate the extent to which hydrocarbons contribute to the formation of O_3 with the aid of chemical mass balance (CMB) receptor modeling. A multivariate statistical technique was also employed to screen ozone scenarios and to classify the study area into several regimes (Yu and Chang 2000).

This study aims to develop methodologies for classifing high-level ozone episodes within the boundary of Seoul City by cluster and disjoint principal component analysis. The high-level ozone episodes in the Seoul area could be divided into common (overall air quality monitoring stations) and site-specific patterns (at a certain station). This study also intends to identify two types of ozone pollution patterns with the application of a fuzzy expert system. Consequently, the high-level ozone episodes, classified in this study, were used as a database for developing an ozone forecasting model using both fuzzy expert and neural network systems (Heo and Kim 2004).

2. DATA

In this study, we used hourly data sets of pollutants and meteorological parameters measured at both surface and upper elevation (500 hPa) stations in Seoul for the period 1989-1996. The air pollutant data was obtained from the Korean Ministry of the Environment (KMOE), and the meteorological data from the Korean Meteorological Administration. Table 1 presents a summary of the data sets used for our model computation. The data contains hourly ozone levels (O_3) along with other pollutants such as: sulfur dioxide (SO_2), carbon monoxide (CO), and nitrogen dioxide (NO_2). The surface meteorological parameters also include wind direction (S-WD), wind speed (S-WS), temperature (S-TEMP), relative humidity (S-RH), and solar radiation (S-SOLAR). In addition, meteorological parameters measured at 500 hPa were provided such as: wind direction (U-WD), wind speed (U-WS), and temperature (U-TEMP). Some portions of the data that were not measured at hourly intervals (i.e., S-TEMP and S-RH (at 3 hr intervals); U-WD, U-WS, U-TEMP (at 12 hr intervals)) were interpolated with a cubic spline into hourly data sets.

The city of Seoul has grown into a teeming metropolis due to rapid urbanization and industrialization over the past 30 years. Nestled among several mountains of varying size, Seoul is situated on the lower reaches of the Han River, which flows through the central part of the Korean Peninsula. These geographical conditions ought be considered among the natural factors affecting Seoul's air quality (Heo and Kim 2002). In this study, four target sites were selected and named: A (Ssangmun), B (Bangi), C (Guro), and D (Gwanghwamun) (Fig. 1). Site selection was determined by geographical factors and previous meteorological records indicating areas where high-level ozone episodes had previously occurred. Station A, set in the northerly region, is surrounded by several mountains that make the area vulnerable to the build-up of pollutants. Station B is located to the southeast, upstream of the Han River, in an area well known for recording frequent episodes of high-level ozone. Station C is located in the western sector, downstream on the Han River in western Seoul. This site was chosen based on it being a semi-industrialized area with many plants and factories that would likely be potential sources for local air pollution. Station D represents a heavy traffic area in the central part of Seoul (Heo and Kim 2004). Detailed information of all air quality monitoring stations investigated in this study has been provided previously (Heo and Kim 2002).

Table 1. A summary of air pollutants and meteorological data for the period 1989-1996 analyzed in this study.

Variable ID	Variable	Units	Interval Given Time
SO_2	Sulfur dioxide	ppb	1 hour
CO	Carbon monoxide	ppb	1 hour
NO_2	Nitrogen dioxide	ppb	1 hour
O_3	Ozone	ppb	1 hour
S-WD	Surface Wind Direction	degree	1 hour
S-WS	Surface Wind Speed	m s^{-1}	1 hour
S-TEMP	Surface Temperature	℃	3 hour
S-RH	Surface Relative Humidity	%	3 hour
S-SOLAR	Surface Solar Radiation	MJ m^{-2}	1 hour
U-WD	500 hPa Wind Direction	degree	12 hour
U-WS	500 hPa Wind Speed	m s^{-1}	12 hour
U-TEMP	500 hPa Temperature	℃	12 hour

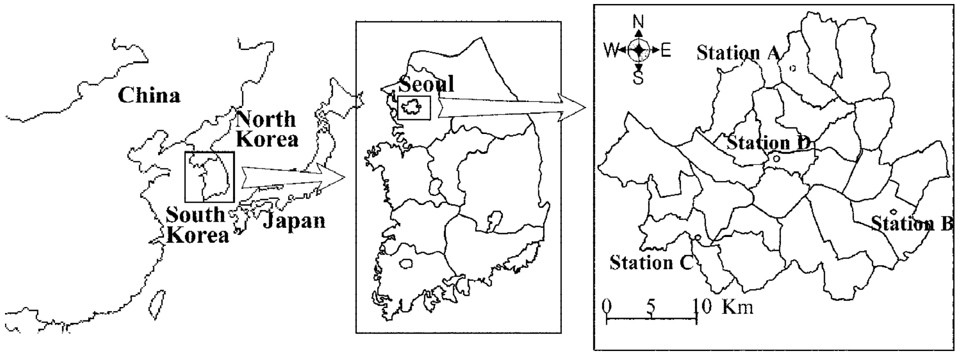

Fig. 1. A geographical location of four air quality monitoring sites in the Seoul area operated by Korea Ministry of Environment.

3. EXPERIMENTAL SECTION

3.1 Data Preparation

At the initial stage of our analysis, all the data sets were sorted out for the application of an ozone forecast model. Since this study aimed at especially forecasting the extent of ozone-level rise, the data sets below 80 ppb of ozone were excluded from further analysis. These screened data sets were then re-divided into two subsets with reference to (1) 80 ~ 99 ppb (80 ppb is the ambient air quality standard for an 8-hour duration by USEPA) and (2) above 100 ppb (100 ppb is the 1-hour based ozone standard by KMOE). These two data groups were then re-divided into two subgroups on the basis of wind direction information at 500 hPa (U-WD). For ozone levels above 80 ppb, 95% of the data was accompanied by a westerly wind (U-WD: 180~360°), for the rest, by an easterly wind (U-WD: 5~175°). Each subset of data was further divided into two subgroups based on surface wind direction (S-WD) between westerly (180~360°) and easterly (5~175°) winds.

By following these procedures, eight sub-data groups (I ~ VIII) were created (Table 2). However in the case of station C, no distinction was made between O_3 concentration levels due to the scarcity of data sets above 100 ppb O_3. If the total quantity of data is compared among the eight data groups, the sum of Groups I and II was dominant with 768 cases (81.4%) out of the 943 ozone episodes. These data groups satisfied simultaneously the meteorological conditions of WD (180~360°) for both the surface and upper elevation levels. In light of the fact that a sufficient amount of data is required for the application of various statistical analyses, only Groups I and II were used in our cluster and disjoint components analysis.

It is difficult to quantify the information associated with wind direction, if expressed as an angle. Hence, the concept of a wind direction index (WD_{index}, Ziomas et al. 1995) was used to yield values in the range of 1 to 3.

$$WD_{index} = 2 + \sin(\psi - \pi / 2), \qquad (1)$$

where ψ is the wind direction expressed as radians. The WD_{index} at both the surface and upper levels was computed for our pattern recognition analysis.

3.2 Pattern Recognition by Cluster and Disjoint Principal Component Analysis

A flow chart for the pattern recognition procedures of ozone episodes in this study is presented in Fig. 2. In our study, the total number of cases of ozone episodes is too large to identify their characteristics without a statistical analysis. Hence, we attempted to identify the characteristics of ozone episodes from large data sets without priori information through cluster and disjoint principal component analysis. In this study, the data sets from Groups I and II were used primarily to exercise pattern recognition. There were a total of 768 cases of ozone episodes in groups I and II (including 216, 283, 82, and 187 cases for stations A through D, respectively). Before grouping similar classes of the ozone episode cases at each station, only 74 cases of above 100 ppb (at station D) were exercised for the optimal classification method; such an approach may effect the correct classification of the data sets.

Table 2. Number of total cases and relative proportion (%) of cases classified after data grouping.

U-WD	180° ~ 360°				5° ~ 175°				Total
S-WD	180°~360°		5° ~ 175°		180°~360°		5° ~ 175°		
Group Site	≥ 100 Group I	80~99 Group II	≥ 100 Group III	80~99 Group IV	≥ 100 Group V	80~99 Group VI	≥ 100 Group VII	80~99 Group VIII	
Station A	74 (27.6)	142 (53.0)	9 (3.4)	21 (7.8)	0 (0.0)	20 (7.5)	2 (0.7)	0 (0.0)	268 (100)
Station B	70 (20.2)	213 (61.4)	2 (0.6)	17 (4.9)	11 (3.2)	27 (7.8)	3 (0.9)	4 (1.2)	347 (100)
Station C	82 (62.6)		19 (14.5)		16 (12.2)		14 (10.7)		131 (100)
Station D	74 (37.6)	113 (57.4)	0 (0.0)	1 (0.5)	5 (2.5)	4 (2.0)	0 (0.0)	0 (0.0)	197 (100)
Sub Total	218 (26.8)	468 (57.6)	11 (1.4)	39 (4.8)	16 (2.0)	51 (6.3)	5 (0.6)	4 (0.5)	812 (100)
Total	768 (81.4)		69 (7.3)		83 (8.8)		23 (2.4)		943 (100)

3.2.1 Application of Cluster Analysis

In the application of cluster analysis, one is required to select a method of computing distance (between cases or objects) and a clustering method to determine similar classes. Hence, distance between the cases was computed on the basis of both the Euclidean and squared Euclidean distance, while average linkage and Ward's methods of hierarchical cluster analysis were used for clustering.

Suppose that x_{ij} (i =1, 2, 3, ...m, j =1, 2, 3, ...n) is the score on the jth measurement variable (SO_2, CO, NO_2, O_3, etc.) for the ith individual case (a case with above 80 ppb ozone). A new value of Z_{ij} is obtained by auto-scaling (z-transformation),

$$Z_{ij} = \left(X_{ij} - \overline{X}_{.j}\right) / \sigma_i \,, \tag{2}$$

where σ_i is the standard deviation for the ith individual case and $\overline{X}_{.j}$ is the mean value of variable j over the whole set of cases. The Euclidean and squared Euclidean distance can be defined as follows:

$$D_{ik} = \sum_{k=1}^{m} \left[\left(Z_{ij} - Z_{kj} \right)^2 \right]^{1/2} , \tag{3}$$

$$D_{ik} = \sum_{k=1}^{m}\left(Z_{ij} - Z_{kj}\right)^2 \quad . \tag{4}$$

It is acknowledged that the Euclidean distance is the simplest measure used in the various research fields (Dubes and Jain 1979). The squared Euclidean distance has been suggested as a reasonable measure for environmental data (Hopke 1985).

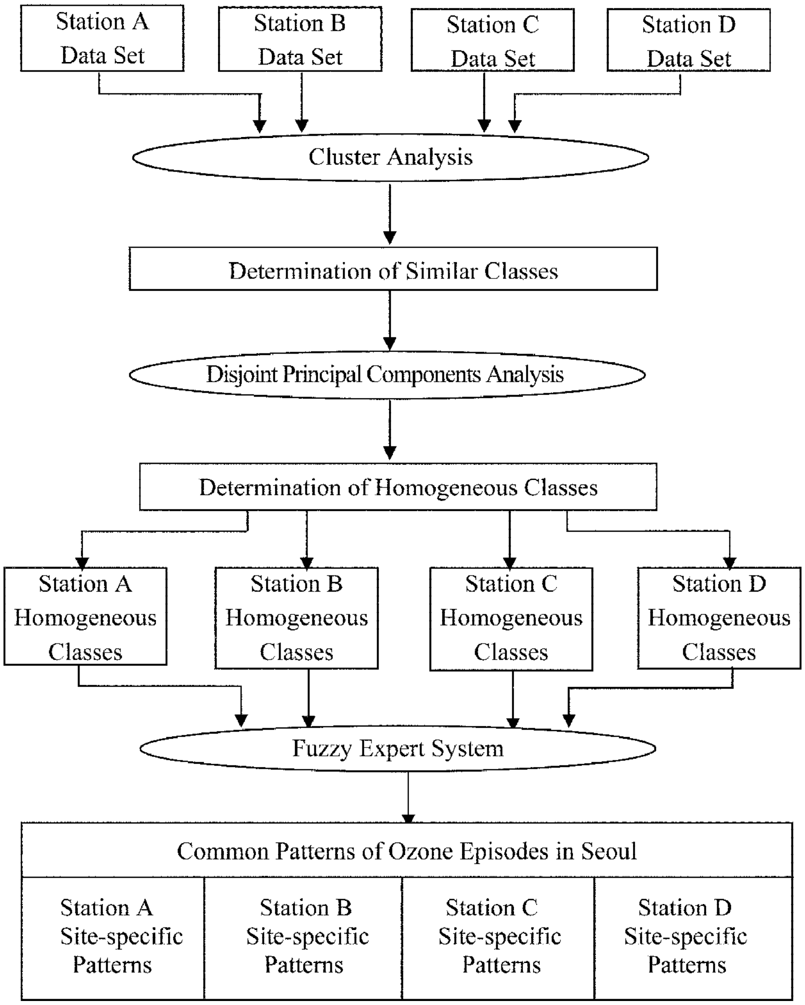
Station A Data Set
Station B Data Set
Station C Data Set
Station D Data Set
Cluster Analysis
Determination of Similar Classes
Disjoint Principal Components Analysis
Determination of Homogeneous Classes
Station A Homogeneous Classes
Station B Homogeneous Classes
Station C Homogeneous Classes
Station D Homogeneous Classes
Fuzzy Expert System
Common Patterns of Ozone Episodes in Seoul
Station A Site-specific Patterns
Station B Site-specific Patterns
Station C Site-specific Patterns
Station D Site-specific Patterns

Fig. 2. A flow chart for the pattern recognition procedures of ozone episodes in this study.

The hierarchical clustering algorithm used is based on the average-linkage and Ward's methods developed for clustering correlation matrices (Hopke 1985). The average linkage method combines the two clusters with the smallest average distance between all pairs of elements in the two clusters. Ward's method combines the two clusters with the smallest increase in the overall sum of squares within cluster distances at each stage. The object of these algorithms is to compute a dendrogram with two-dimensional diagrams; they can be used to illustrate the fusions (or divisions) made at each successive stage of the analysis to help assemble all elements into a single tree. The dendrogram in Fig. 3 was obtained on the basis of the average linkage method and Euclidean distance. Large groups were assigned at quite low levels if clustering and small cases appeared near the end of the dendrogram (as shown on the right side of Fig. 3). Cluster analysis, on the basis of research experience, has a disadvantage in determining a cut-off distance (Hopke 1985). In Fig. 3, only four major classes were chosen based on a certain degree of dissimilarity. Thus, nine cases (marked as stars in the dendrogram near the case ID) in several minor classes were defined as outliers.

3.2.2 Application of SIMCA (Soft Independent Modeling of Class Analogy)

Principal component analysis (PCA) has been applied to identify the dominant multivariate relationships in the measured data. To investigate quantitatively how well classes were assigned by cluster analysis, disjoint principal components analysis was conducted by the soft independent modeling of class analogy [SIMCA, Wold (1976)]. The SIMCA is a statistical package for the analysis of pattern recognition; it is based on fitting the principal components (PC) models that can facilitate thes classification of all data sets. To this end, the input data sets were divided into either a training or test set. The former is a class known either by prior information or by statistical analysis (such as cluster analysis), while the latter consists of unassigned cases. The PC models are hence used to designate the unknown cases (test set) into a single or several known classes (training set). Consequently, if some cases are unclassified, they can be treated as outliers. The SIMCA can be used to treat the data as two-step processes: (1) the development of PC models using a training set, and (2) the classification of cases in a test set according to their degree of fit to the different PC models.

The most important process of PCA is to determine the number of significant components. In the SIMCA, these numbers are determined by the cross validation criteria of Wold (1978). In this study, if the total number of cases is small for one class (i.e., less than 6), the significant components cannot be determined. This class will then be discarded (from building the PC model) as an outlier (i.e., class 99). A distance of a case can be calculated for each PC model and compared to a critical distance (at a 95% probability level). The PC models can then be compared with the assigned degree for each case by a standard decision plot. Homogeneous classes were formed through comparison between the PC models; cases that did not belong to any class were thus classified as outliers. However, the number of those outliers can be reduced by the optimal classification method.

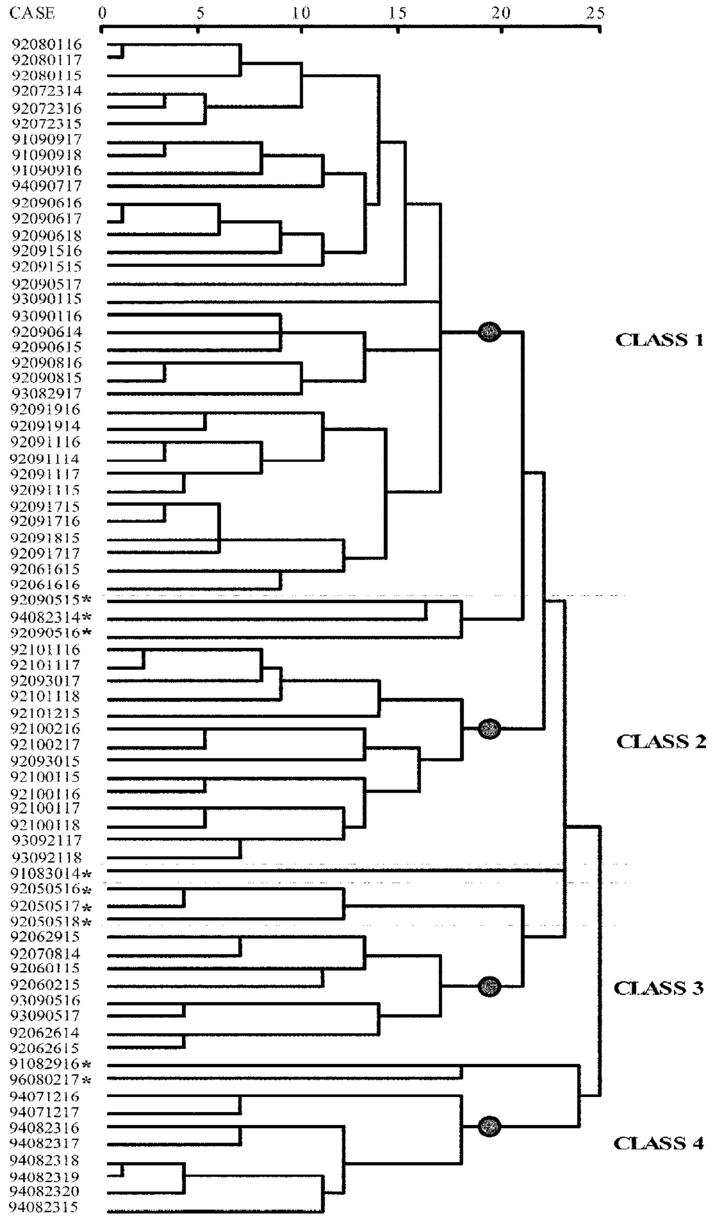

Fig. 3. A dendrogram based on the average linkage method and the Euclidean distance for 74 cases of above 100 ppb ozone at Station D.

3.2.3 Determination of the Optimal Classification Method

At station D, only 74 cases of above 100 ppb ozone were analyzed to determine the optimal classification method of cluster, distance measurement, and transformation methods. Figure 4 presents a procedure of several statistical analyses executed in this study for a correct classification.

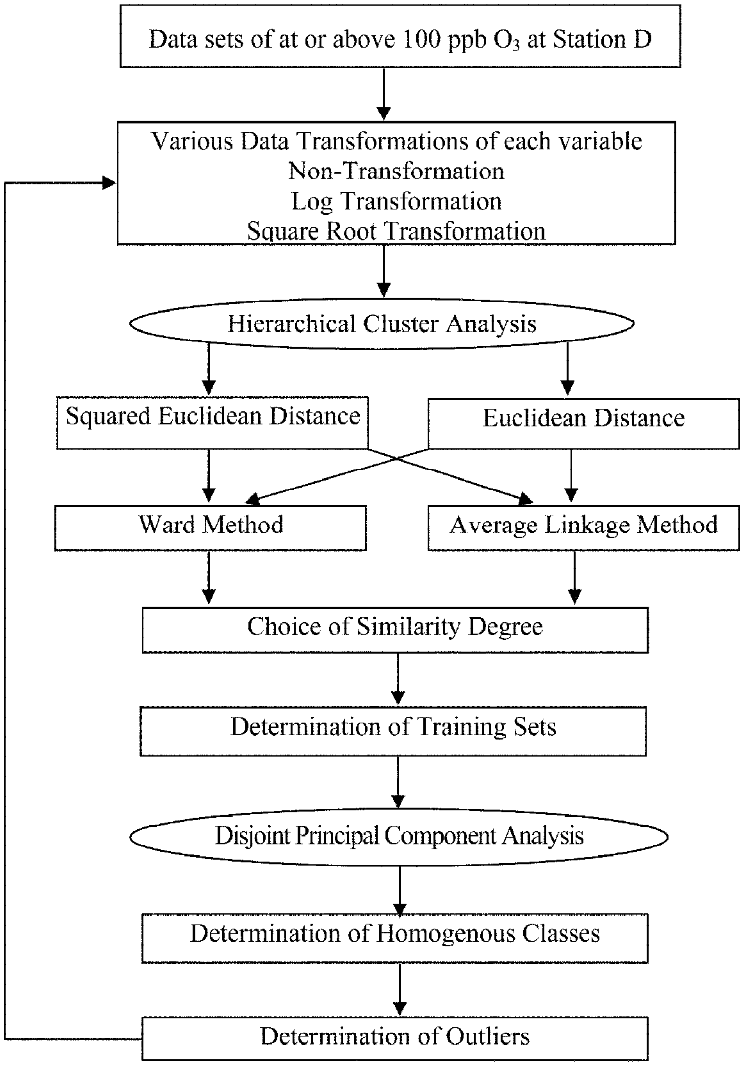
Data sets of at or above 100 ppb O_3 at Station D
Various Data Transformations of each variable
Non-Transformation
Log Transformation
Square Root Transformation
Hierarchical Cluster Analysis
Squared Euclidean Distance
Euclidean Distance
Ward Method
Average Linkage Method
Choice of Similarity Degree
Determination of Training Sets
Disjoint Principal Component Analysis
Determination of Homogenous Classes
Determination of Outliers

Fig. 4. A flow chart of the procedures to determine the optimal classification method in the pattern recognition analysis.

In many cases, the use of skewed distribution for each variable can yield an incorrect classification of data sets; hence the data sets need to be processed by proper transformation techniques (e.g., the use of natural logarithm and square root) (Vong et. al. 1985). Thus, each variable was tested via three different types (e.g., natural logarithm, square root, and non-transformation). To determine the optimal classification method, one variable (e.g., SO_2) was transformated into three different types. Then, cluster analysis was performed in concert with two clustering methods and two similarity measures. Following this, similar cases were assigned to one of the classes based on degree of similarity. Each class assigned was then used as a training set for the disjoint principal component analaysis. Homogeneous classes were determined through comparison between the PC models. When the number of outliers reached minimum value, the classification method applied at the first step was taken to be a pertinent one. This procedure was also carried out iteratively for the accompanying variables (refer to Table 1) in the same manner. All cases identified initially as class 99 were reprocessed until they were finally determined as outliers. Thus, class 99 was a dumping area for outlier cases or a tentative reservoir to allocate new episodic cases.

Optimal clustering was obtained both by the average linkage cluster method with Euclidean distance and by a number of different methods including logarithmic transformation (CO and U-WS), square root transformation (S-TEMP), and non-transformation (SO_2, NO_2, O_3, S-RH, S-WD_{index}, S-WS, S-SOLAR, U-TEMP, and U-WD_{index}). The other three stations were also analyzed by the same procedure.

3.3 Pattern Recognition by the Fuzzy Expert System

Many patterns of high-level ozone episodes at the four stations were classified using statistical analyses presented in section 3.2. The ozone episodes in the Seoul area could be divided into common and site-specific patterns. The common patterns could explain some ozone episodes at the right target stations as well as other stations in Seoul. To classify common and specific patterns, an ozone episode pattern at one station was compared with another by use of the fuzzy expert system of analysis (as shown Fig. 2). In addition, to sort out high-level ozone episode cases from the massive data sets, ozone episode pattern was expressed as a rule by the fuzzy expert system.

3.3.1 Fuzzy Expert System

Fuzzy logic has been extended to handle the concept of partial truth, a truth-value between ‘completely true’ and ‘completely false’. Fuzzy logic may contain some elements with a partial degree of membership. If X is a collection of objects denoted by x, then a fuzzy set $\tilde{A}$ in X can be expressed as a set of ordered pairs:

$$\tilde{A} = \left\{ \left(x,\ \mu_{\tilde{A}}(x) \right) | x \in X \right\}. \qquad (5)$$

Where $\mu_{\tilde{A}}(x)$ is called the membership function or the degree of truth for x in $\tilde{A}$. The fuzzy expert system is the most common use of fuzzy logic. In general, the fuzzy expert system consists of (1) the knowledge base, (2) fuzzification, (3) inference, and (4) defuzzification. The knowledge base is equivalent to the database including the linguistic variables, the fuzzy sets, and the fuzzy production rules. The fuzzification is to determine the extent of compatibility that input data has to each appropriate fuzzy set via membership functions. For the inference process, the given facts (i.e., the membership values for the input variable) are analyzed and new statements are derived (i.e., the rule conclusion). In the inference process, a given rule is evaluated using three steps: aggregation, implication, and accumulation. The defuzzification allows conversion of a fuzzy output set (a result of the inference process) into a crisp value (Zimmermann 1990). In this study, a fuzzy expert system, included in the DataEngine software package (MIT 1997), was used for this purpose.

3.3.2 Database and Membership Function

The homogeneous classes for groups I and II and the other data groups (III, IV, V, VI, VII, and VIII), as classified in section 3.1, were used as a database for a sub-rule base to build a unified fuzzy rule base for each station (Table 3). Of all 13 classes in station A, three classes of data group I and eight classes of data group II were classified by pattern recognition analysis. Station B was assigned a total of 24 classes; in this case, 47 out of 347 cases were classified as class 99 outliers. The data sets of station C were classified into a total of eight classes with 10 out of 131 cases characterized as class 99 outliers. For station D, four classes of data group I, and eight classes of data group II were determined. A total of 39 out of 197 cases were newly grouped as class 99 outliers.

A data set of a certain class can be used to provide a rule base for the fuzzy model. Thus, a binding of rules from each class can provide a unified rule. Unified rules for each station can then be named as Rule Stations' A through D. For example, Rule Station A, consisting of 13 sub-rules (or homogeneous classes), has 199 cases; Likewise, Rule Stations B, C, and D have 24, 8, and 13 sub-rules, respectively.

To build a membership function of input variables for each class, this study used the simplest membership functions consisting of both triangular and trapezoidal types with straight lines. Whereas the latter uses four points to make an input membership function, the former is made up of three points for an output membership function. The input membership functions can hence be explained by basic statistical information of 12 variables in each class. The membership function of all input variables for each class is defined by the following four points:

(1) the difference between a minimum value and a 5% value in the total range of each variable in each class: $x1$,

(2) the difference between an average value and a standard deviation of each variable in each class: $x2$,

(3) the sum of an average value and a standard deviation of each variable in each class:$x3$,

(4) the sum of a maximum value and a 5% value in the total range of each variable in each class: $x4$.

The fuzzy set of variables j in class k includes $[x_1, \mu(x_1)]$, $[x_2, \mu(x_2)]$, $[x_3, \mu(x_3)]$, and

$[x_4, \mu(x_4)]$. Here, $\mu(x_i)$ is the membership function of x_i. In addition, a decision rule was made to consist of a premise (IF...) linked by a conclusion (THEN...) which was based on multiple input and single output structures. A value of output variable in each class was calculated by the membership functions and the decision rules determined. For inference procedure in this study, a minimum operator was used at the aggregation step, a product operator at the implication step, and a maximum operator at the accumulation step. A centroid method was used in the defuzzification procedure.

Table 3. An array of each homogeneous class classified by cluster analysis and disjoint principal components analysis. All classes for each station are used as database to build a unified rule via the fuzzy expert system. Class ID is used to designate different classes for a given station. [No. of Cases (Class ID)]

Class	Station A	Station B	Station C	Station D
1	17 (1GroupI-A)	22 (1GroupI-B)	32 (1GroupI&II-C)	31 (1GroupI-D)
2	20 (2GroupI-A)	7 (2GroupI-B)	8 (2GroupI&II-C)	13 (2GroupI-D)
3	10 (3GroupI-A)	6 (3GroupI-B)	23 (3GroupI&II-C)	7 (3GroupI-D)
4	17 (1GroupII-A)	7 (4GroupI-B)	5 (4GroupI&II-C)	8 (4GroupI-D)
5	20 (2GroupII-A)	6 (5GroupI-B)	4 (5GroupI&II-C)	21 (1GroupII-D)
6	4 (3GroupII-A)	13 (6GroupI-B)	19 (GroupIII&IV-C)	15 (2GroupII-D)
7	13 (4GroupII-A)	6 (1GroupII-B)	16 (GroupV&VI-C)	16 (3GroupII-D)
8	22 (5GroupII-A)	10 (2GroupII-B)	14(GroupVII&VIII-C)	18 (4GroupII-D)
9	16 (6GroupII-A)	6 (3GroupII-B)	–	6 (5GroupII-D)
10	5 (7GroupII-A)	10 (4GroupII-B)	–	6 (6GroupII-D)
11	5 (8GroupII-A)	25 (5GroupII-B)	–	4 (7GroupII-D)
12	30 (GroupIII&IV-A)	24 (6GroupII-B)	–	4 (8GroupII-D)
13	20 (GroupV&VI-A)	13 (7GroupII-B)	–	9 (GroupV&VI-D)
14	–	5 (8GroupII-B)	–	–
15	–	10 (9GroupII-B)	–	–
16	–	7 (10GroupII-B)	–	–
17	–	19 (11GroupII-B)	–	–
18	–	16 (12GroupII-B)	–	–
19	–	4 (13GroupII-B)	–	–
20	–	5 (14GroupII-B)	–	–
21	–	15 (15GroupII-B)	–	–
22	–	19 (GroupIII&IV-B)	–	–
23	–	38 (GroupV&VI-B)	–	–
24	–	7 (GroupVII&VIII-B)	–	–
99	69 (Class 99) (including GroupVII)	47 (Class 99)	10 (Class 99)	39 (Class 99) (including GroupIV)
Total	268	347	131	197

* Groups I~VIII were classified by the data grouping scheme given in Table 2.

3.3.3 Reclassification of Each Homogeneous Class by the Fuzzy Expert System

For each station, homogeneous classes were reclassified by comparing fuzzy rule bases. In this phase, the data sets for total cases were viewed in respect of the unified rule which is a combination of sub-rules at each station. For example, if any sub-rule in the unified rule included specific cases for station D, then the sub-rules were classified as Specific Rule Station D. If any sub-rule applied to cases at three or more stations, it was determined as a Common Rule, common patterns for Seoul. Finally, the common and site-specific rules for the four target stations (i.e., Common Rule; Specific Rule Station A, Specific Rule Station B, Specific Rule Station C, and Specific Rule Station D) were determined by the above procedure.

4. RESULTS AND DISCUSSION

4.1 Results of Classification

Table 4 presents the final structure of the Common Rule which consists of eight sub-rules (or patterns) with a total of 331 cases. Class 1 with a total of 38 cases in the Common Rule was made up of 2 (Station A), 5 (Station B), 6 (Station C), and 25 cases (Station D). It was found that Class 2 with 15 cases (4.5%) was deficient of the data from station C. Class 7 did not include any data for station D. The Common Rule consisted of (1) 118 (44%) out of 268 (Station A); (2) 106 (30.5%) out of 347 (Station B); (3) 54 (41.2%) out of 131 (Station C); and (4) 53 (26.9%) out of 197 cases (Station D).

Judging from the data structure of the Common Rule, the common patterns of ozone episodes in the Seoul area can be explained by 44% of ozone episodes at station A. Most ozone episodes at station D can be explained by site-specific patterns. This result is similar to the previous study of Ghim and Chang (2000) who reported that the distribution of ozone concentrations at station D was affected most sensitively by photochemical reactions associated with local emissions. The Common Rule comprised 331 cases (35.1%) out of 943 that exhibited ozone concentrations above 80 ppb at the four stations.

Table 5 presents the structures of both common and site-specific rules classified by the fuzzy expert system. For example, Specific Rule Station A had 11 classes, including a total of 150 cases; 10 of those cases were assigned to class 99. In Specific Rule Station B with 20 classes, the patterns of ozone episodes appeared to be the most diverse. Specific Rule Station C was made up of 8 classes. In Specific Rule Station D, 10 classes were identified with a total of 144 cases.

4.2 Analyses of Ozone Episodes

In an empirical forecasting model, it is important to collect many past patterns of ozone episodes (databases) to improve forecast accuracy. A forecasting model based on common and unique patterns can be used flexibly to solve the site-specificity problem of ozone pollution and to extend databases for high forecast accuracy.

Various ozone episode patterns classified by the fuzzy expert system were analyzed in

Table 4. Comparison of the whole data structure in the system Common Rule.

Class	Station A	Station B	Station C	Station D	Total
1	2 (0.6%*)	5 (1.5%)	6 (1.8%)	25 (7.8%)	38 (11.5%)
2	9 (2.7%)	3 (0.9%)	0 (0.0%)	3 (0.9%)	15 (4.5%)
3	9 (2.7%)	19 (5.7%)	3 (0.9%)	2 (0.6%)	33 (10.0%)
4	9 (2.7%)	6 (1.8%)	4 (1.2%)	3 (0.9%)	22 (6.7%)
5	56 (16.9%)	29 (8.8%)	16 (4.8%)	8 (2.4%)	109 (32.9%)
6	7 (2.1%)	4 (1.2%)	2 (0.6%)	10 (3.0%)	23 (7.0%)
7	9 (2.7%)	9 (2.7%)	14 (4.2%)	0 (0.0%)	32 (9.7%)
8	17 (5.1%)	31 (9.4%)	9 (2.7%)	2 (0.6%)	59 (17.8%)
Total	118(35.7%)	106(32.0%)	54(16.3%)	53 (16.0%)	331 (100%)

* Parenthesis denotes a portion of cases in each class to a total of 331 cases.

terms of: 1) the basic statistics of each variable in each class; and 2) the mean ratios of each variable between each class and the whole 943 cases.

4.2.1 Common Patterns for Ozone Episodes in Seoul

Figure 5 presents the characteristics of eight different patterns in the Common Rule. This type of approach allows comparative analysis of the trends for all variables at a time. For Class 1, temperature and wind speed in the upper atmosphere (500 hPa) averaged -14.6°C and 19.6 m s^{-1}, respectively. The ozone episode of Class 1 may be explained mainly by the transport of primary pollutants under strong wind in the upper layer. In Class 2, ozone levels might have been raised by high NO_2 concentrations (50.1 ppb on an average) under weak wind speed in the upper layer (6.8 m s^{-1}). For Class 4, most cases were observed in the summer between 16:00 and 19:00 in spite of weak solar radiation (5.0 MJ m^{-3} on an average). In this case, ozone concentrations exceeding 80 ppb were maintained until early evening. The surface weather conditions of high temperatures (30.8°C on an average) and weak wind speeds (2.6 m s^{-1}) under low NO_2 levels might have been conducive to the enhancement of ozone (e.g., Fuentes and Dann 1994).

It was found that four Classes (3, 5, 7, and 8) arose under conditions of strong solar radiation and high temperatures at the surface. In Class 3, 78.8% of cases belonged to the late spring months of May and June, while 21.2% of cases could be attributed to the remainder of summer. Class 5 was dominated by cases with high temperatures at both the surface (30.8°C on an average) and upper layer (-5.4°C) which might have helped elevate ozone levels. High ozone concentrations in Class 7 could be ascribed to the stagnation of primary pollutants under high surface temperatures but weak surface winds from the east. Because the presence of

Table 5. The case numbers and their fraction (%) for each class in the Common Rule and the site-specific rules classified by the fuzzy expert system.

Class (Sub-rule)	Common Rule	Specific Rule-Station A	Specific Rule-Station B	Specific Rule-Station C	Specific Rule-Station D
1	38 (11.5%)	7 (4.7%)	5 (2.1%)	8 (10.4%)	10 (6.9%)
2	15 (4.5%)	5 (3.3%)	6 (2.5%)	4 (5.2%)	7 (4.9%)
3	33 (10.0%)	4 (2.7%)	7 (2.9%)	7 (9.1%)	7 (4.9%)
4	22 (6.6%)	22 (14.7%)	6 (2.5%)	4 (5.2%)	4 (2.8%)
5	109 (32.9%)	4 (2.7%)	8 (3.3%)	19 (24.7%)	4 (2.8%)
6	23 (7.0%)	19 (12.7%)	9 (3.7%)	16 (20.8%)	17 (11.8%)
7	32 (9.7%)	16 (10.7%)	9 (3.7%)	7 (9.1%)	21 (14.6%)
8	59 (17.8%)	17 (11.3%)	19 (7.9%)	4 (5.2%)	17 (11.8%)
9	-	13 (8.7%)	4 (1.7%)	-	40 (27.8%)
10	-	19 (12.7%)	10 (4.1%)	-	5 (3.5%)
11	-	14 (9.3%)	6 (2.5%)	-	-
12	-	-	15 (6.2%)	-	-
13	-	-	19 (7.9%)	-	-
14	-	-	5 (2.1%)	-	-
15	-	-	44 (18.3%)	-	-
16	-	-	19 (7.9%)	-	-
17	-	-	6 (2.5%)	-	-
18	-	-	10 (4.1%)	-	-
19	-	-	8 (3.3%)	-	-
20	-	-	7 (2.9%)	-	-
99	-	10 (6.7%)	19 (7.9%)	8 (10.4%)	12 (8.3%)
Total	331 (100%)	150 (100%)	241 (100%)	77 (100%)	144 (100%)

primary pollutants (such as NO_2, CO, and SO_2) is a prerequisite for ozone formation, their concentrations reduced noticeably during the ozone peaks. Class 8 was characterized as the period during which the easterly wind was dominant in the upper layer. The common patterns for ozone episodes in Seoul generally showed a strong similarity in that the accumulation of abundant primary pollutants might have played a key role in the formation of high daytime ozone levels when solar radiation and temperatures were highest and wind speeds low.

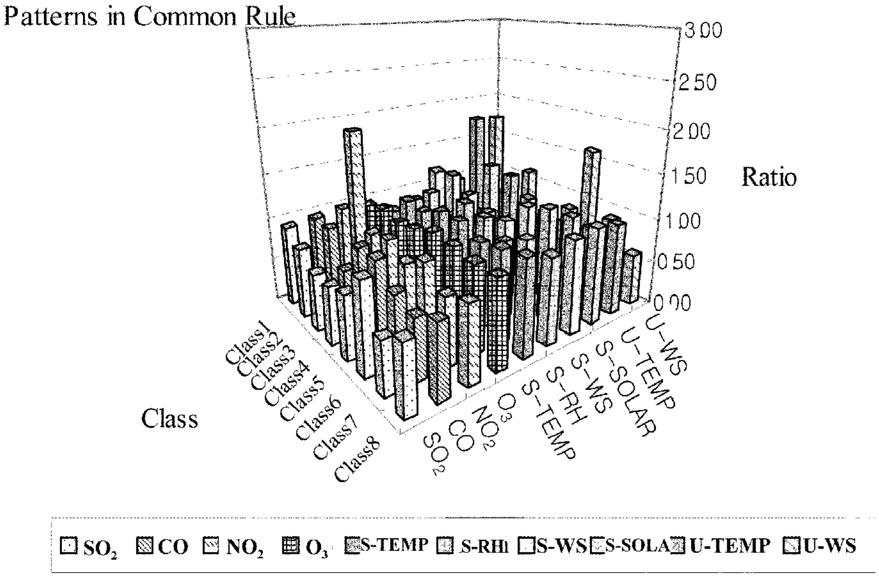

Fig. 5. The common eight patterns of ozone episodes in the Seoul area. The ratios were calculated by means of each variable in each class divided by those in all cases with ozone levels above 80 ppb.

4.2.2 Site-specific Patterns of Ozone Episodes

The basic features of each class in each specific rule at the four stations are presented in Figures 6a, b, c and d. For Specific Rule Station A, high levels of ozone in Class 5 were seen persistently until the evening accompanied by weak winds and high temperatures. For Class 11, high levels of ozone occurred continuously during the day (12:00, July 23 to 15:00, July 25, 1994). This class was characterized by high temperatures and weak wind speed at both the surface and the upper level . This class may represent a pattern in which high levels of ozone build up continuously under a sufficient supply of NOx in the urban atmospheres at night due to insufficient scavenging (Bower et al. 1994; Fuentes and Dann 1994).

For Specific Rule Station B, Class 4 was characterized by low RH (34%, average) and low temperatures at both levels (19.0 °C at the surface to -17.5 °C at the upper level). In Class 9, high levels of ozone were accompanied by high NO_2 (36.0 to 64.0 ppb) and low RH (24.0 to 54.0%). The meteorological conditions in Class 20 were characterized by weak winds (1.3 m s^{-1} at the surface mainly easterly and 4.8 m s^{-1} at the upper level) and low RH (37.8%) during the day.

For Specific Rule Station C, the cases in Class 1 occurred during the early dawn of July 3 and 4, 1992. In this class, high levels of ozone formed during the day were sustained until late at night under high RH (92.8%) and calm winds (0.9 m s^{-1}). Liu et al. (1990) observed the

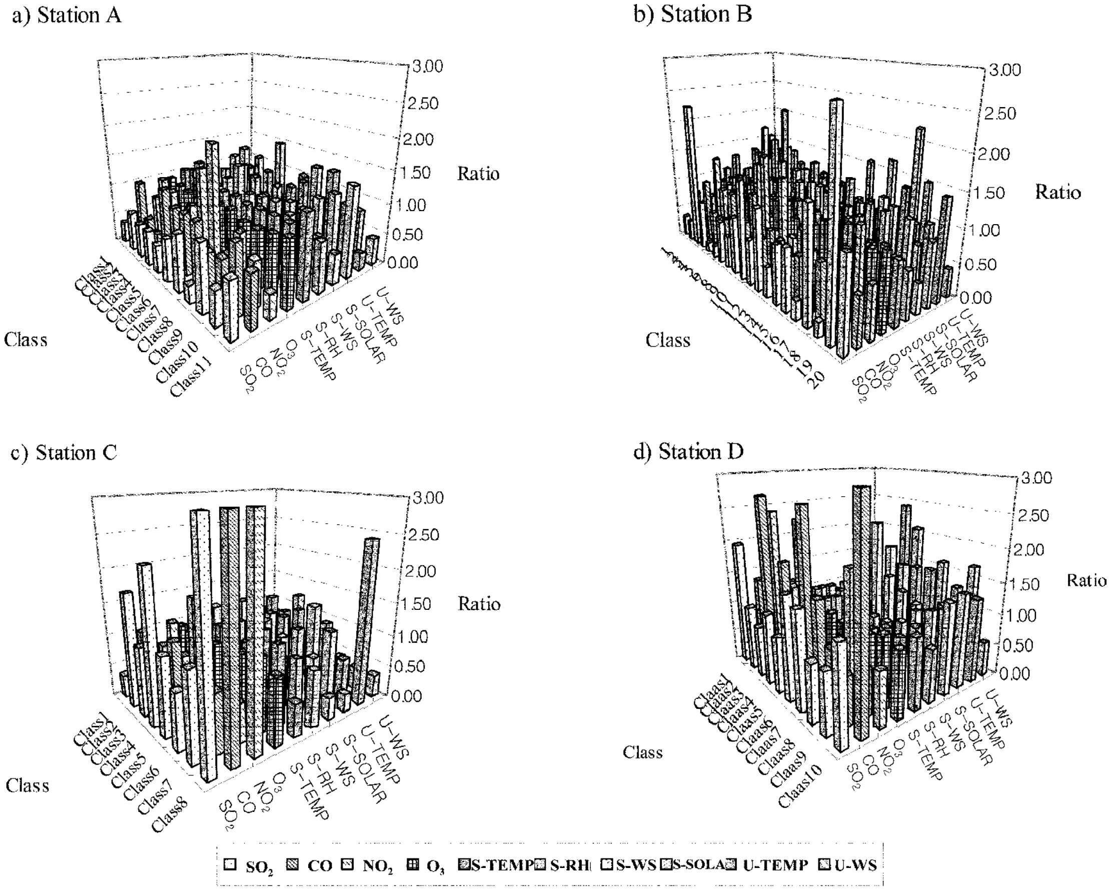

Fig. 6. The specific patterns of ozone episodes at Stations A, B, C, and D.

presence of a second ozone peak during the early dawn in Taipei; they explained the cause of such phenomena in terms of meteorological condition (e.g., weak wind speeds) rather than photochemical production. Class 8 was represented by high levels of ozone (4 April, 1994) under low temperatures both at the surface and upper level. High levels of ozone during this period might have occurred due to the accumulation of primary pollutants (SO_2: 227-251 ppb, CO: 61-92 ppb, NO_2: 63-131 ppb) at weak winds. In this case, the strength of influential variables was characterized in the order: SO_2 > CO > NO_2 > O_3 > S-RH > S-TEMP > U-TEMP > U-WS > S-WS > S-SOLAR.

For Specific Rule Station D, Class 1 consisted of the data sets obtained during 1 to 2 October, 1992; this appeared to be a common pattern for the fall-term ozone episodes. High ozone concentrations of this class were found to be affected by relatively high pollutant levels (including CO, SO_2, and NO_2 concentrations), in spite of low surface temperatures (22.9 °C) and solar radiation (7 MJ m^{-3}). Thus, high levels of ozone in this class might have occurred mainly because of the precursor accumulation under weak surface winds. In Class 4, high ozone concentrations lasted until the late evening (18:00 to 20:00); this phenomena may be accounted for by the prevailing high temperatures (30.7 °C at the surface to -4.4 °C at the upper level) in concert with weak winds (2.4 m s^{-1} at the surface and 5.6 m s^{-1} at the upper level). Class 5 was made up of the data between 13:00 and 16:00 on 5 May, 1994. In this class, the strength of influential variables was found in the order: S-WS > U-WS > S-SOLAR > NO_2 > U-TEMP > SO_2 > O_3 > S-RH > S-TEMP > CO. In addition, low RH (37.8%) RH > S-TEMP > CO. In addition, low RH (37.8%) also might have been conducive to the enhancement of ozone. The patterns in Specific Rule Station D showed enhanced values of primary pollutants, solar radiation, and wind speed compared to their counterparts in the Common Rule. When the diurnal variations of ozone levels were examined at station D, the ozone levels from late night to early morning of the next day were much lower than those at the other stations in Seoul. Thus, the occurrance of high-level ozone episodes at this station might have resulted from the typical photochemical production and destruction of ozone in NO_x concentrations under heavy traffic conditions (Heo and Kim 2002).

5. CONCLUSIONS

In this study, multivariate statistical methods (such as cluster and disjoint principal component analysis) were employed to classify past ozone episodes in Seoul, Korea (during the period of 1989-1996) with an aid of a fuzzy expert system. In our pattern recognition analysis, the optimal clustering was determined for up to 12 variables by the employment of an average linkage cluster method with Euclidean distance after transformation of all the relevant environmental parameters. In addition, by means of the fuzzy expert system, the common ozone episode cases were classified into eight different patterns. The site-specific ozone episode cases were classified as follows: 11 patterns (Station A), 20 (Station B), 8 (Station C), and 10 (Station D).

The approaches used in this study were aimed to develope methodologies for the following objectives: (1) solving problems objectively on the subjectively interpreted cluster analysis; (2) explaining non-linear relationships among various variables affecting ozone formation; and (3) collecting useful information from complex pollution phenomena. We also intended to provide some perspectives in which fuzzy expert systems can be used effectively in distinguishing ozone episode cases from massive data sets. One of the most valuable results in this study was to describe the characteristics of high-level ozone using long-term pollution and weather data sets.

The selection of the patterns for diverse ozone episodes is a principal step in the designing stage of an ozone forecast model. In our companion work, we presented a new ozone forecasting model which was built based on common and site-specific ozone patterns identified in this study (Heo and Kim 2004). The model based on fuzzy expert and neural network systems was developed to predict daily maximum 1-h ozone concentrations in Seoul. In terms of the performance of this model, it can select a past ozone episode pattern suitable for input data and can forecast the daily maximum ozone concentrations by a neural network model based on this pattern; thus, the more available ozone episode data, the more accurate forecast. The accuracy of our ozone forecast model could be continuously improved by adding more reliable and recent data sets.

Acknowledgements This work was partly supported by the Korea Research Foundation (99-005-E-00025) as a part of the study for abatement of malodorous VOC/VIC.

REFERENCES

Ballester, E. B., G. C. Valls, J. L. Carrasco-Rodriguez, E. S. Olivas, and S. Valle-Tascon, 2002: Effective 1-day ahead prediction of hourly surface ozone concentrations in eastern Spain using linear models and neural networks. *Ecolog. Model.*, **156**, 27-41.

Bower, J. S., K. J. Stevenson, G. F. J. Broughton, J. E. Lampert, B. P. Sweeney, and J. Wilken, 1994: Assessing recent surface ozone concentrations in the U.K.. *Atmos. Environ.*, **28**D, 115-128.

Chaloulakou, A., M. Saisana, and N. Spyrellis, 2003: Comparative assessment of neural networks and regression models for forecasting summertime ozone, Athens. *Sci. Total Environ.*, **313**, 1-13.

Chan, L. Y., C. Y. Chan, and Y. Qin, 1998a: Surface ozone pattern in Hong Kong. *Am. Meteorol. Soc.,* October, 1153-1165.

Chan, L. Y., H. Y. Liu, K. S. Lam, T. Wang , S. J. Oltmans, and J. M. Harris, 1998b: Analysis of the seasonal behavior of tropospheric ozone at Hong Kong. *Atmos. Environ.,* **32**, 159-168.

Chen, J., S. Islam, and P. Biswas, 1998: Nonlinear dynamics of hourly ozone concentrations: Nonparametric short term prediction. *Atmos. Environ.*, **32**, 1839-1848.

Chung, J., R. A. Wadden, and P. A. Scheff, 1996: Development of ozone-precursor relationships using VOC receptor modeling. *Atmos. Environ.*, **30**, 3167-3179.

Comrie, A. C., 1997: Comparing neural networks and regression models for ozone forecasting. *J. Air Waste Management Assoc.*, **47**, 653-663.

Dorling, S. R., T. D. Davies, and C. E. Pierce, 1992: Cluster analysis: A technique for estimating the synoptic meteorological controls on air and prediction chemistry - Method and applications. *Atmos. Environ.*, **14**, 2575-2581.

Dubes, R., and A. K. Jain, 1979: Clustering techniques- the users dilemma. *Pattern Recognition*, **8**, 247-260.

Gardner, M. W., and S. R. Dorling, 2000: Statistical surface ozone models: an improved methodology to account for non-linear behavior. *Atmos. Environ.*, **34**, 21-34.

Ghim, Y. S., and Y. S. Chang, 2000: Characteristics of ground-level ozone distributions in Korea for the period of 1990-1995. *J. Geophys. Res.*, **105**, 8877-8890.

Fuentes, J. D., and T. F. Dann, 1994: Ground-level ozone in Eastern Canada: seasonal variation, trends, and occurrences of high concentrations. *J. Air Waste Management Assoc.*, **44**, 1019-1026.

Hadjiiski, L., and P. Hopke, 2000: Application of artificial neural networks to modeling and prediction of ambient ozone concentrations. *J. Air Waste Management Assoc.*, **50**, 894-901.

Hanna, S. R., G. E. Moore, and M. E. Fernau, 1996: Evaluation of photochemical grid models (UAM-IV, UAM-V, and the ROM/UAM-IV couple) using data from the lake Michigan ozone study (LMOS). *Atmos. Environ.*, **30**, 3265-3279.

Heo, J. S., and D. S. Kim, 2002: The characterization of surface ozone concentrations in Seoul, Korea. *J. Korean Soc. Atmos. Environ.*, **18**(E3), 129-142.

Heo, J. S., and D. S. Kim, 2004: A new method of ozone forecasting using fuzzy expert and neural network systems. *Sci. Total Environ.*, **325**, 221-237.

Hopke, P. K., 1985: Receptor medeling in environmental chemistry, *John Wiley & Sons Inc., New York,* 210-224 pp.

Huang, G., 1992: A stepwise cluster analysis method for predicting air quality in an urban environment. *Atmos. Environ.*, **3**, 349-357.

Hubbard, M. C., and W. G. Cobourn, 1998: Development of a regression model to forecast ground-level ozone concentration in Louisville, K.Y.. *Atmos. Environ.*, **32**, 2637-2647.

Liu, C. M., S. C. Liu, and S. H. Shen, 1990: A study of Taipei ozone problem. *Atmos. Environ.,* **6**, 1461-1472.

Lu, W. Z., W. J. Wang, H. Y. Fan, A. Y. T. Leung, Z. B. Xu, S. M. Lo, and J. C. K. Wong, 2002: Prediction of pollutant levels in Causeway bay area of Hong Kong using an improved neural network model. *J. Environ. Eng.*, December, 1146-1157.

MIT, 1997: Management intelligenter technologien GmbH DataEngine. Tutorials and Theory, 1st Edition. Aachen, Germany.

Peton, N., G. Dray, M. Mesbah, and B. Vuillot, 2000: Modelling and analysis of ozone episodes. *Environ. Model. Software*, **15**, 647-652.

Prybutok, V. R., J. Yi, and D. Mitchell, 2000: Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations. *European J. Operational Res.*, **122**, 31-40.

Robeson, S. M., and D. G. Steyn, 1990: Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations. *Atmos. Environ.,* **2**, 303-312.

Ryan, W. F., 1995: Forecasting severe ozone episodes in the Baltimore Metropolitan area. *Atmos. Environ.*, **29**, 2387-2398.

Startis, J. A., R. D. Tsitouridou, and V. D. Simeonov, 1995: Chemometrical classification of aerosol analytical data. *Toxicolog. Environ. Chem.*, **47**, 191-196.

Vong, R. J., I. E. Frank, R. J. Charison, and B. R. Kowlaski, 1985: Exploratory data analysis of rainwater composition. In: Breen J.J., (Eds). Environmental application of chemometrics. *Am. Chem. Soc.*.

Vogt, W., 1987: Cluster analysis in clinical chemistry: a model. *John Wiley & Sons, Inc., New York,* 12-62 pp.

Vukovich, F. M., 1995: Regional-scale boundary layer ozone variations in the Estern United States and their association with meteorological variations. *Atmos. Environ.*, **29**, 2259-2273.

Wang, X., W. Lu, W. Wang, and A. Y. T. Leung, 2003a: A study variation trend within area of affecting human health in Hong Kong. *Chemosphere*, **52**, 1405-1410.

Wang, W., W. Lu, X. Wang, and A. Y. T. Leung, 2003b: Prediction of maximum daily ozone level using combined neural network and statistical characteristics. *Environ. Inter.*, **29**, 555-562.

Wold, S., 1976: Pattern recognition by means of disjoint principal component models. *Pattern Recognition*, **8**, 27-139.

Wold, S., 1978: Cross-validation estimation of the number of components in factor and principal components model. *Technometrics*, **20**, 397-405.

Yoo, S. J., and D. S. Kim, 1997: Classification of ambient particulate samples using cluster analysis and disjoint principal component analysis. *J. Korea Air Pollution Res. Assoc.,* **13**, 51-63 (in Korean).

Yu, T. Y., and L. F. W. Chang, 2000: Selection of the scenarios of ozone pollution at southern Taiwan area utilizing principal component analysis. *Atmos. Environ.*, **34**, 4499-4509.

Zimmermann, H. J., 1990: Fuzzy set theory - and its applications. *Kluwer Academic Publishers. Norwell, Massachusetts.*

Ziomas, I. C., D. Meals, C. S. Zerefos, A. F. Bais, and A. G. Paliatsos, 1995: Forecasting peak pollutant levels from meteorological variables. *Atmos. Environ.*, **29**, 3703-3711.