Journal of Hazardous Materials 183 (2010) 448–459

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

Predictive monitoring and diagnosis of periodic air pollution in a subway station

YongSu Kim[a], MinJung Kim[a], JungJin Lim[a], Jeong Tai Kim[b,**], ChangKyoo Yoo[a,*]

[a] Department of Environmental Science and Engineering, Center for Environmental Studies/Green Energy Center, Kyung Hee University, 1 Seochon-dong, Giheung-gu, Yongin-si, Gyeonggi-do 446-701, Republic of Korea
[b] Department of Architectural Engineering, Center for Sustainable Healthy Buildings, Kyung Hee University, 1 Seochon-dong, Giheung-gu, Yongin-si, Gyeonggi-do 446-701, Republic of Korea

ARTICLE INFO

Article history:
Received 3 March 2010
Received in revised form 6 July 2010
Accepted 10 July 2010
Available online 16 July 2010

Keywords:
Air quality monitoring
Lifted model
Multiway principal component analysis (MPCA)
Periodic pattern
Predictive monitoring

ABSTRACT

The purpose of this study was to develop a predictive monitoring and diagnosis system for the air pollutants in a subway system using a lifting technique with a multiway principal component analysis (MPCA) which monitors the periodic patterns of the air pollutants and diagnoses the sources of the contamination. The basic purpose of this lifting technique was to capture the multivariate and periodic characteristics of all of the indoor air samples collected during each day. These characteristics could then be used to improve the handling of strong periodic fluctuations in the air quality environment in subway systems and will allow important changes in the indoor air quality to be quickly detected. The predictive monitoring approach was applied to a real indoor air quality dataset collected by telemonitoring systems (TMS) that indicated some periodic variations in the air pollutants and multivariate relationships between the measured variables. Two monitoring models – global and seasonal – were developed to study climate change in Korea. The proposed predictive monitoring method using the lifted model resulted in fewer false alarms and missed faults due to non-stationary behavior than that were experienced with the conventional methods. This method could be used to identify the contributions of various pollution sources.

1. Introduction

Advanced monitoring and control strategies for atmospheric environments are attracting renewed interest due to increasingly stringent environmental regulations, because concerns about the effects of the air quality of indoor microenvironments on public health are increasing. The Korea Ministry of Environment (MOE) established the indoor air quality (IAQ) act in an attempt to control major pollutants including PM_{10}, CO_2, CO, VOC and formaldehyde in indoor environments such as subway platforms. There is a strong interest in monitoring air quality to quickly detect and identify any fault or abnormality that might negatively affect the air quality, because of the increasingly stringent air quality requirements imposed by law [1].

Many variables in the systems process or environment are recorded on-line or off-line in modern systems, and the number of variables recorded is increasing due to the development of new electronic sensors. Proper techniques are required to extract useful information from the extensive amount of recorded data [2]. Subway station sites in a metro are considered to be air quality "hot spots," which also include heavily trafficked roadways and power plants. Indoor air quality in subway stations can be strongly influenced by one pollution source, particularly if the station is located downwind of the source [3]. Therefore, a major objective of the monitoring system is to quickly detect the occurrence of assignable causes of contaminated air quality so that detailed measurements and steps to correct ventilation issues can be undertaken in order to improve the indoor air quality. The current status in a system should be monitored in order to meet all of the operational targets for quality, safety constraints and environmental constraints at a minimum cost.

In metro systems, traditional monitoring systems have been based on time-series analysis which measures and monitors a single particulate matter pollutant (PM_{10} and/or $PM_{2.5}$), which is also known as univariate monitoring. Univariate monitoring charts are widely used to monitor a small number of key pollutant variables in an air pollution monitoring system that is capable of detecting the occurrence of any event having a special or assignable cause. However, monitoring only a single or a few variables is not adequate, because many variables are correlated and interrelated and therefore have an effect on one another. More specifically, univariate diagnostics is not sufficient for detecting highly concentrated

* Corresponding author. Tel.: +82 31 201 3824; fax: +82 31 202 8854.
** Corresponding author. Tel: +82 31 201 2539; fax: +82 31 206 2109.
E-mail addresses: jtkim@khu.ac.kr (J.T. Kim), ckyoo@khu.ac.kr, ChangKyoo.Yoo@biomath.ugent.be (C. Yoo).

0304-3894/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2010.07.045

pollutants or outliers, because the environmental measurements are not independent. Univariate methods may result in the masking of underlying factors [4]. Therefore, a monitoring system that takes these correlations into consideration, a multivariate monitoring method, should be used for adequate control of indoor air quality management. Multivariate monitoring that examines all of the air pollutants simultaneously can determine how all of the pollutants are behaving relative to one another. This is particularly useful for examining pollution, because several pollutants are often present at the same time and can have simultaneous effects on human health [5].

It is important to understand the dynamics and phenomena of the environmental process being studied in order to monitor and interpret the generic status of indoor air quality. In other words, improvements in the monitoring process can be achieved by obtaining better knowledge of the system by answering the following questions: which variables characterize the process, what are their internal interactions and what degrees of confidence can be attributed to these measurements? All of these questions are related to the characterization of the system, which involves several fundamental stages: the description of the system, the listing of the variables that characterize the system behavior, the establishment of models between the variables, the identification of the parameters which intervene in these models, the simplification of the models to make them compatible with real-time use, and the validation of these models [6]. Multivariate monitoring methods, including principal component analysis (PCA) and partial least squares (PLS), have been applied to monitor the air quality in many environmental systems [1,7–9] [1,3,7–9]. In 2009, Lin [1] used a multivariate time-series model to simulate the PM_{10} concentrations by studying the potential effects that meteorological factors and co-pollutants may have on the day-to-day variations in PM_{10} concentrations and to predict the PM_{10} peaks. They applied the two statistical methods of principal component analysis (PCA) and cluster analysis (CA) to the NO_2 and PM_{10} concentrations obtained from the air quality monitoring network in the city and found that the variations in NO_2 and PM_{10} concentrations exhibited patterns similar to the variations in traffic volume.

The indoor air quality in most subway stations is subject to large periodic fluctuations due to fluctuations in the number of passengers, the number of trains and the ambient air conditions. Since the concentrations of the air pollutants tend to fluctuate widely over a day, their means and variances do not remain constant with time. Therefore, the use of monitoring methods that utilize conventional, multivariate, statistical processes that implicitly assume a stationary, underlying environmental system may lead to many false alarms and missed faults. Better monitoring performance can be achieved by accounting for this periodic pattern when applying a new monitoring process to this system.

In this study, a predictive monitoring and diagnosis method for indoor air quality in a subway station was developed to tackle both the problems associated with the multivariate correlation of the indoor air pollutants and the non-stationary problem of periodic changes in the concentrations of the pollutants. The use of a multiway, unfolding concept for a given periodic property can improve the overall monitoring performance for all of the air pollutants.

This paper is organized as follows. First, the data set used in this study, data from a telemonitoring system (TMS) in a subway line in Korea, is explained. Second, the multiway principal component analysis (MPCA) is briefly discussed to show the basic concepts involved in monitoring a batch process and a cyclic continuous process. Third, a predictive monitoring and diagnosis method is recommended in order to detect the occurrence of any event having a special or assignable cause. The Section 3 presents an illustrative application, and then the conclusions are presented in Section 4.

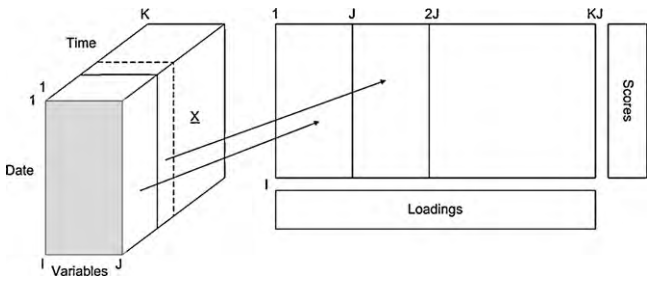

Fig. 1. Unfolding of the three-dimensional process data [12].

2. Material and methods

2.1. Data from a telemonitoring system (TMS)

In this study, air pollutant data sets from telemonitoring systems (TMSs) in four Korean subway stations were used. TMSs were built for the management of indoor air quality in subways. The TMSs were located at the center of the subway platforms and measured the concentration levels of seven air pollutants. The proposed data sets consisted of NO, NO_2, NO_X, PM_{10}, $PM_{2.5}$, CO and CO_2 concentrations on a time-scale of one sample per hour (hourly averaged), plus the two meteorological parameters of temperature and humidity, in a subway station. The NO, NO_2, and NO_X concentrations were measured using the chemiluminescences of nitro-oxide materials and ozone. The PM_{10} and $PM_{2.5}$ concentrations were measured using the beta-ray attenuation principle with corresponding size distribution filters. The CO and CO_2 concentrations were measured by studying the non-dispersive infrared (IR) radiation absorption by carbon monoxide and carbon dioxide molecules at specific wavelengths. The meteorological conditions of temperature and relative humidity strongly influenced the efficiency of the photochemical processes leading to a noisy measurement and the formation of the subsidiary particulates PM_{10} and $PM_{2.5}$ [10,11].

2.2. Multiway principal component analysis (MPCA)

MPCA is an expanded method of PCA, which can be used to monitor and analyze a batch process. MPCA can compress the data and extract the information by projecting the data into a low-dimensional space that summarizes both the variables and their time histories during normal operation. The new data is then monitored by comparing the progress of the projections of its variable trajectories in the reduced space with the statistical distribution of the trajectories from past normal operations [2,12,13]. Batch data is typically reported in terms of batch numbers, variables and times. The data is arranged into a three-dimensional matrix $\underline{\mathbf{X}}$ ($I \times J \times K$), where I is the number of batches, J is the number of variables and K is the number of times each batch is sampled. This matrix can be decomposed using various three-way techniques, one of which is MPCA. MPCA is equivalent to performing ordinary PCA on a large two-dimensional matrix $\mathbf{X}$ constructed by unfolding the three-way data in the manner shown schematically in Fig. 1 [14,15].

MPCA decomposes the three-way array $\underline{\mathbf{X}}$ into a summation of the product of a score t_r and a loading matrix P_r plus a residual array $\underline{\mathbf{E}}$ that is minimized in the least squares sense as follows:

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} t_r \otimes P_r + \underline{\mathbf{E}} = \sum_{r=1}^{R} t_r p_r^T + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E}, \quad (1)$$

where $\otimes$ denotes the Kronecker product ($\underline{\mathbf{X}} = \mathbf{t} \otimes \mathbf{P}$ is $\underline{\mathbf{X}}(i, j, k) = t(i)\mathbf{P}(j, k)$), R denotes the number of principal components retained, t_r expresses the relationship among batches, p_r is related to the variables and their time variation, and $\mathbf{E}$ is the residual matrix. The first expression in equation (1) represents the 3-D decomposition, while the second expression corresponds to the more common 2-D decomposition [12].

The statistics used for monitoring multivariable batch processes are Hotelling's T^2-statistic and the Q-statistic [12,16]. The T^2-statistic is a Mahalanobis distance between the new data and the center of the normal operating condition data in a reduced dimension. The pattern of the residuals is monitored using the Q-statistic, which is also referred to as the squared prediction error (SPE). The T^2-statistic monitors systematic variations in the principal component (PC) subspace, while the Q-statistic represents variations not explained by the retained PCs. In other words, faults in the process that violate the normal correlation of variables are detected in the PC subspace by the T^2-statistic, whereas faults that violate the PCA models are detected in the residual space by the Q-statistic. At the end-of-batch, the T^2- and Q-statistics for batch i are calculated as follows:

$$T_i^2 = \mathbf{t}_r{}^T \mathbf{S}^{-1} \mathbf{t}_r \sim \frac{R(I^2-1)}{I(I-R)} F_{R,I-R-1} \text{ and} \tag{2}$$

$$Q_i = \mathbf{e}_i \mathbf{e}_i^T = \sum_{c=1}^{KJ} \mathbf{E}(i,c)^2, \tag{3}$$

where $\mathbf{e}_i$ is the ith row of $\mathbf{E}$, I is the number of batches in the reference set, $\mathbf{t}_r$ is a vector of R scores, $\mathbf{S}$ is the $(R \times R)$ covariance matrix of the t-scores calculated during the model development (which is diagonal due to the orthogonality of the t-score values), R is the number of PCs retained in the model, and $F_{R,I\text{-}R\text{-}1}$ is the F-distribution value with R and I–R–1 degrees of freedom [17]. The statistical limits on the T^2- and Q-statistics are computed by assuming that the data conforms to a multivariate normal distribution. The confidence limits of the T^2-statistic are calculated from the F-distribution. The distribution of the Q-statistic is calculated from the Chi-squared distribution, $\text{SPE}_{k,\alpha} = (\nu_k/2m_k)\chi^2_{2m_k^2/\nu_k,\alpha}$, where m_k and ν_k are the mean and variance of the *SPE*, respectively, and $\chi^2_{2m_k^2/\nu,\alpha_k}$ is the critical value of the χ^2 variable with $2m_k^2/\nu_k$ degrees of freedom at significance level of α [16,17].

We considered a current batch to be in-control if it had a $100(1-\alpha)$ % confidence for a new sample, $\mathbf{x}_{\text{new}}$, if $T^2_{\text{new}} < T^2_{\text{lim}}$ and $Q_{\text{new}} < Q^2_{\text{lim}}$. Otherwise, a batch was designated as out of control. Here, the T^2 value was used to detect faults associated with abnormal variations within a model subspace, whereas the Q value was used to detect new events that were not taken into account in the model subspace.

We outline an MPCA-based method for real-time monitoring of the progress of batch or periodic processes, such as indoor air quality monitoring. An indoor air quality is monitored in the reduced space defined by the principal components of the MPCA model. The loading matrices from the MPCA of the reference database contain most of the structural information describing how the variable measurements deviate from their average trajectories under normal operation [12]. The details of the procedure are as follows.

(A) *Develop normal operating condition (NOC) model*

(1) Get data and unfold $\underline{\mathbf{X}}(I \times J \times K)$ to $\mathbf{X}(I \times JK)$.
(2) The data $\mathbf{X}(I \times JK)$ are normalized using the mean and standard deviation of each variable at each time in the batch cycle over all batches.
(3) Apply PCA to the scaled data and obtain the score matrix $\mathbf{T}(I \times R)$ and loading matrix $\mathbf{P}(JK \times R)$, where R is the number of principal components.
(4) For each batch of $\mathbf{X}(I \times JK)$, $\mathbf{X}^T(I \times JK)$ is considered and projected into the loading space. Scores and residuals are calculated from $\mathbf{t}^T(1 \times R) = \mathbf{x}^T\mathbf{P}$ and $\mathbf{e}(JK \times 1) = \mathbf{x} - \mathbf{Pt}$. For a total of I batches, $\mathbf{e}$ constitutes the residual matrix $\mathbf{E}(I \times JK)$.
(5) Calculate T^2, SPE_k and obtain their confidence limits.

(B) *On-line monitoring*

(1) For new batch data recorded up to time k, $\mathbf{X}_{\text{new}}(k \times J)$, unfold it to $\mathbf{x}^T_{\text{new}}(1 \times Jk)$. Apply the same scaling as was used in the modeling.
(2) For the scaled $\mathbf{x}^T_{\text{new}}(1 \times Jk)$, fill in the missing values to generate $\mathbf{x}^T_{\text{new}}(1 \times Jk)$ using one of the three filling approaches to fill the future data. Then, calculate $\mathbf{t}_{\text{new}}$ and $\mathbf{e}_{\text{new}}$: $\mathbf{e}_{\text{new}}$: $\mathbf{t}^T_{\text{new}}(1 \times R) = \mathbf{x}^T_{\text{new}}\mathbf{P}$, $\mathbf{e}_{\text{new}}(JK \times 1) = \mathbf{x}_{\text{new}} - \mathbf{Pt}_{\text{new}}$
(3) Calculate T^2 and SPE_k

$$T^2 = \mathbf{t}^T_{\text{new}} \mathbf{S}^{-1} \mathbf{t}_{\text{new}}$$

$$\text{SPE}_k = \sum_{c=(k-1)J+1}^{Jk} \mathbf{e}_{\text{new}}(c)^2$$

where $\mathbf{e}_{\text{new}}(c)$ is cth element of $\mathbf{e}_{\text{new}}$.
(4) Determine whether T^2 or SPE_k exceeds its confidence limit.

2.3. Predictive cyclic monitoring of indoor air pollutants using MPCA

Univariate monitoring methods are often difficult to use when important faults or events occur in indoor air quality, because the signal to noise ratio in the measurement system of air pollutants is lower than the ratio in chemical measurement systems. Multivariate monitoring methods that can extract key information can treat all of the air pollutant data simultaneously and analyze how all of the air pollutants affect one another. Time-series data of air pollutants in a subway station can be expressed as a three-dimensional matrix ($\underline{\mathbf{X}}(I \times J \times K)$). In 1 day, $j = 1, 2, \ldots, J$ variables are measured at $k = 1, 2, \ldots, K$ time intervals, and similar air pollution profiles are collected over several days ($i = 1, 2, \ldots, I$). The purpose of MPCA is to unfold this matrix in order to obtain a two-dimensional matrix on which PCA can then be performed.

The schemes of the predictive monitoring and the diagnosis of the indoor air pollutant concentrations with the periodicity are shown in Fig. 2.

First, the air pollutant data was collected from a telemonitoring system in a subway station and was checked for outliers. Second, off-line and on-line models of MPCA were constructed with the complete, normal daily data in order to monitor the current status of the indoor air pollutants within the period (day) and in a period-to-period (day-to-day) time frame, and both global and local seasonal models were constructed. The pollutant trajectories and the variable relationships between the air pollutants could be interpreted using the scores and loadings of the MPCA model. The on-line MPCA model was used in the subway station to monitor the status of the air quality during the day-to-day time frame and also within the day. In other words, the on-line MPCA model was used to monitor the current status of the air pollutants at the platform of the subway station (the sample time of the sensors was 1 h). After detecting an abnormal status, a contribution plot was used to identify the sources of the contamination.

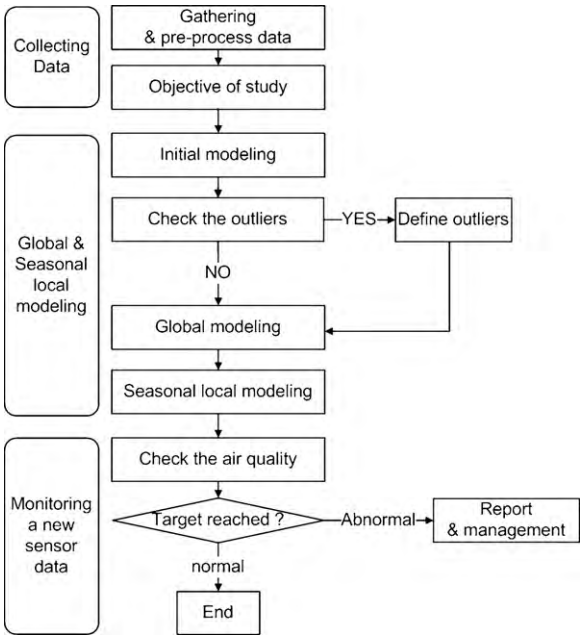

Fig. 2. The scheme of the predictive monitoring and the diagnosis of indoor air pollutant concentrations with the periodicity.

The results of the present study are important, because the monitoring performance was improved by the development of a multiway prediction method for the periodic properties of the indoor air pollutants. The unfolding capability of the MPCA to analyze the predictive time histories of the air pollutants is very useful for the periodicity of this kind of environmental system, because the data from periodic air pollutants usually varies with time. An improvement in the results can be expected by explicitly accounting for the periodic patterns of the air pollutants while applying advanced monitoring and control strategies to the air pollution management system.

3. Results and discussion

3.1. Monitoring system

In this study, seven air pollutants: NO, NO_2, NO_X, PM_{10}, $PM_{2.5}$, CO and CO_2, and two meteorological variables (temperature and humidity) were measured by TMSs in four subway stations. The data was collected between February 2007 and July 2008 with sampling intervals of 1 h. In this study, the air pollutant concentrations in the subway station exhibited seasonal patterns depending on the change in climate in Korea. In order to determine the local seasonal variations in the pollutants over four seasons, the data was classified into four groups, where each of the seasonal groups was used to monitor the characteristics of the air pollutants during one of the four seasons. In this study, MATLAB's PLS toolbox and SIMCA^{P+} software were used to analyze the data sets.

Fig. 3 shows the univariate monitoring charts for the air pollutant concentrations at the S-monitoring station in Korea. These include the concentration profiles of NO, NO_2, NO_X, PM_{10} and $PM_{2.5}$ in the normalized scale [0–1]. The concentrations of some of the pollutants were either high or low compared to the environmental limits determined by the MOE, but this univariate monitoring chart did not take into consideration the correlations between air pollutants. As shown in Fig. 3a, the trends between NO and NO_X and between PM_{10} and $PM_{2.5}$ confirmed the correlation between these pollutants.

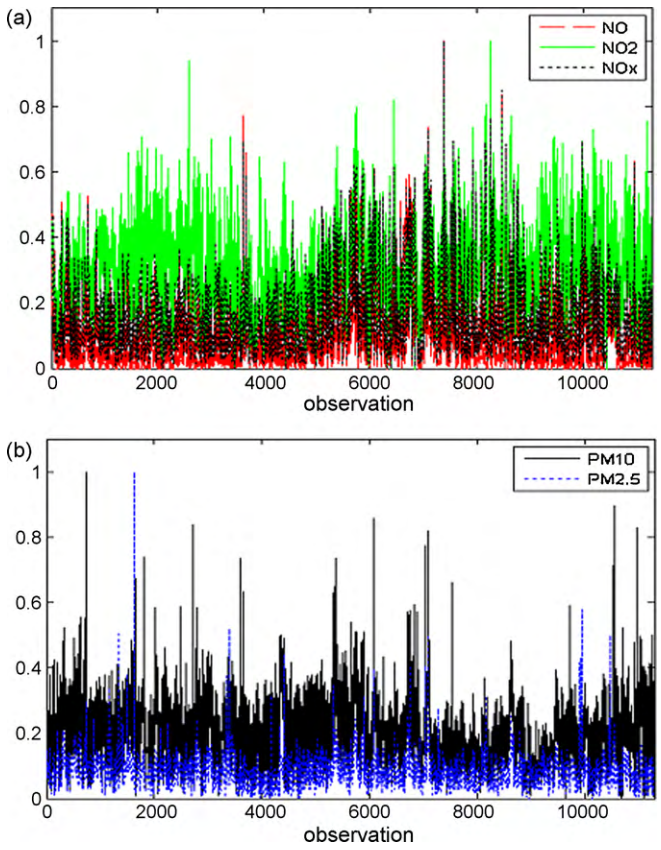

Fig. 3. The univariate monitoring charts for the air pollutant concentrations at a monitoring station: (a) NO, NO_2, NO_X and (b) PM_{10}, $PM_{2.5}$.

Fig. 4 presents the diurnal and weekly variations in the pollutant concentrations and the number of passengers, where the UGM unit of PM_{10} and $PM_{2.5}$ is $\mu g/m^3$. The strong correlation between the number of passengers and the hourly average concentrations of CO_2 and PM_{10} at the platform and the ticket gate in the subway station can be seen in Fig. 4.

It was observed that the concentration of each of the pollutants had a very similar profile (the peaks and valley points were similar) depending on the number of passengers. It was also observed that the CO_2 and PM_{10} concentrations simultaneously increased twice during rush hour, which indicated that the increase in the number of passengers caused the increase in the emission source of the pollutants. There were good correlations evident between CO_2 and PM_{10}, because these pollutants originated from similar sources. In addition, the concentrations of other pollutants including CO, $PM_{2.5}$, NO, NO_2 and NO_X were also highly influenced by the number of passengers during the entire period (not shown in the figure). The fact that the concentrations were lower during the weekend than on weekdays was referred to as the weekend effect. It was also evident in Fig. 4 that the main pollution source in the subway stations was the passengers, and multivariate pollutant variables had strong correlations between them. The correlation between the pollutants should be considered for air pollutant monitoring [11].

Fig. 5 shows the hourly variations in PM_{10}, $PM_{2.5}$ and NO_X concentrations between the four seasons during 1 year. It is evident in Fig. 5 that the patterns of the seasons were similar, and the concentration levels varied slightly among the seasons. Fig. 5 also indicates that there was a large difference between the daily minimum and maximum concentrations throughout all of the seasons. The concentrations of PM_{10}, $PM_{2.5}$ and NO_X started to increase in

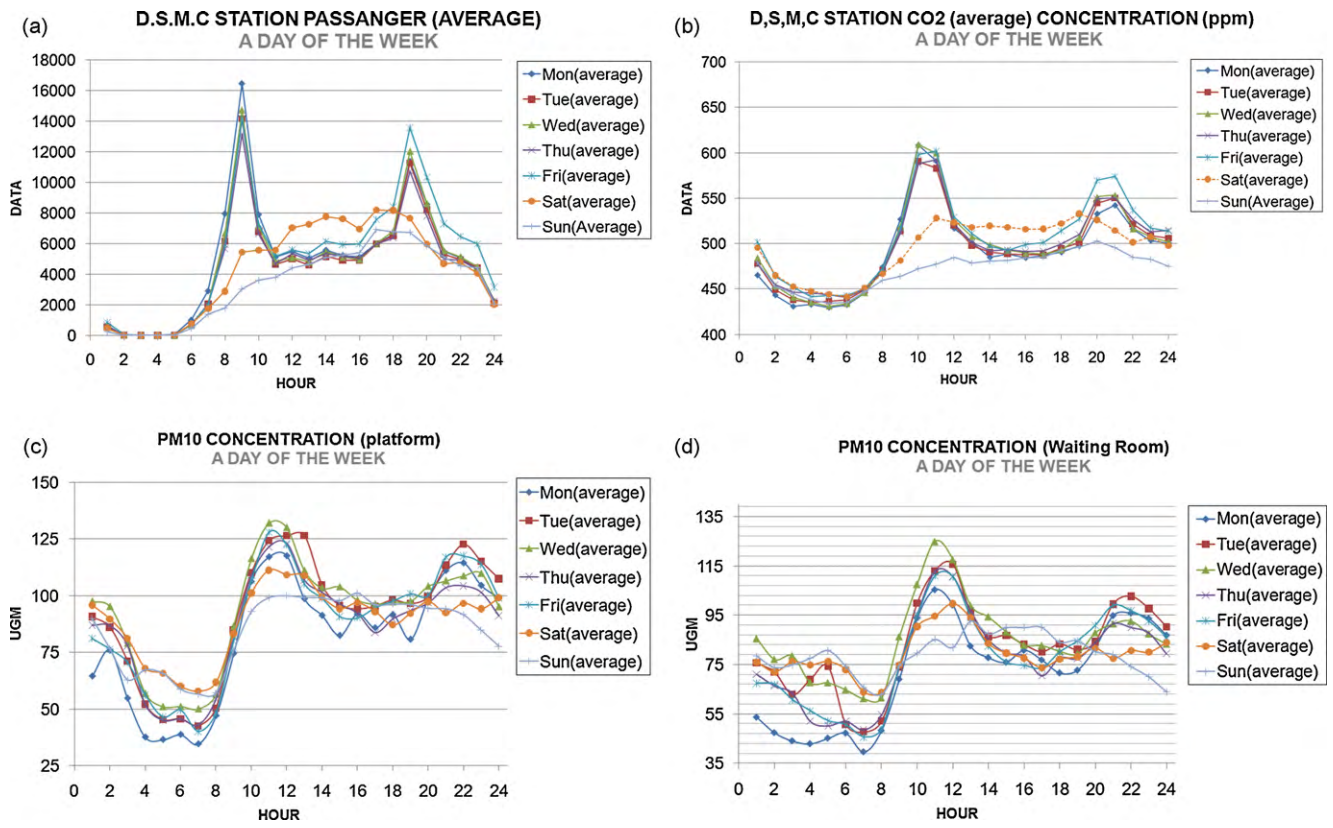

Fig. 4. The diurnal variations for 1 week in the correlation between the number of passengers and the concentrations of CO_2 and PM_{10} on the platforms and at the ticketing locations in four subway stations: (a) number of the passengers, (b) CO_2 concentration, (c) PM_{10} at platform, and (d) PM_{10} at ticketing place (Kim et al., 2009).

the morning hours, reached a peak at noon and then gradually decreased until a second peak at night.

In Fig. 5(a), the daily variability in pollutant concentrations was greater in spring and autumn than in summer and winter. The relative difference between the daily maximum and minimum concentrations tended to be larger during spring and autumn than during summer and winter, indicating that seasonal local sources such as yellow-storm events have a strong influence on particulate matter. Fig. 5(b) shows peaks in the $PM_{2.5}$ concentrations during all of the seasons at noon and again during the late evening, while the daily minimum concentrations were observed in the early morning hours. The $PM_{2.5}$ concentrations remained at the highest levels during the spring. These results also indicate that seasonal local sources have a strong influence on the $PM_{2.5}$ concentrations. Fig. 5(c) shows that the NO_X concentration exhibited a distinct concentration pattern during all of the seasons. The NO_X concentration increased during the daytime before reaching peak levels and then gradually decreased during the nighttime. Fig. 5(a) and (b) also shows a distinct pattern in the NO_X concentration which remained high during the winter season, while the PM concentrations were higher in the spring than during the other seasons. These results indicate that the use of the heating system increased during the winter season which influenced the rate of fuel use and therefore the NO_X concentration.

Figs. 4 and 5 presented useful information including the hourly and weekly variations of each of the pollutant concentrations. During the entire 1-week period, the concentration of each of the pollutants remained at a higher level during the daytime and then gradually decreased at night. In addition, the concentrations were higher during the weekends than they were on weekdays during all of the seasons. As noted above, this phenomenon was influenced by the rate of heating fuel use and/or the number of passengers. These results provided data for the adequate control of ventilation systems in a metro. In addition, a seasonal variation was observed in the concentrations of each of the pollutants, and different patterns were observed for each of the pollutants. These results indicate that interrelationships between those pollutants that exhibited similar patterns should be taken into consideration, and an appropriate pollution control method should be based on the season.

3.2. Predictive monitoring and interpretation

Since univariate monitoring and multi-scatter plots are not suitable for showing relationships between objects and variables, the

Table 1
The conventional PCA and periodic monitoring model results.

PC number	Conventional PCA				A periodic monitoring			
	Global model		Seasonal model		Global model		Seasonal model	
	Eigen value	Cumulative variance	Eigen value	Cumulative variance	Eigen value	Cumulative variance	Eigen value	Cumulative variance
PC 1	3.68	0.408	337	0.374	18.4	0.383	15.7	0.327
PC 2	2.22	0.655	195	0.591	1.0	0.592	12.2	0.576
PC 3	0.989	0.765	161	0.77	3.86	0.672	4.75	0.67

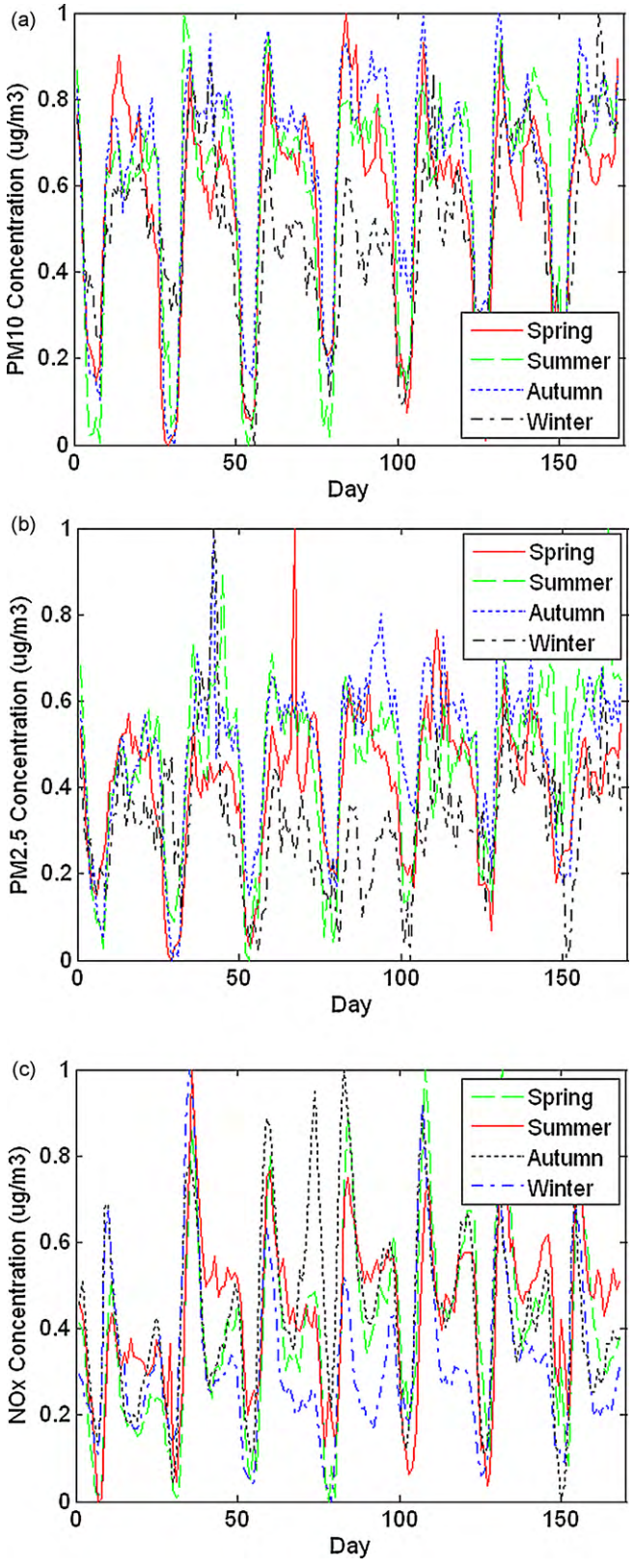

Fig. 5. The hourly average plots of the air pollutant concentrations during four seasons: (a) PM_{10}, (b) $PM_{2.5}$ and (c) NO_X.

conventional multivariate statistical method, PCA, was applied to fully study the seven air pollutants. However, periodic characteristics should be considered due to the dynamic characteristics of indoor air pollutants over time.

One year's worth of data (March 2007 through February 2008) was used for the training period. A subset of normal operational data without significant disturbances was selected in order to develop the training models. It is important to determine the number of principal components (PC) for the PCA model. The number of PCs should be determined when taking into consideration both the dimensionality and the loss of information. Several techniques exist for determining the number of PCs, but there is no single dominant technique [18]. The screen plot and cross-validation criterion were used in this study.

Using a conventional PCA and periodic monitoring, a global model for monitoring an entire data set was identified, and the results are shown in Table 1.

A periodic monitoring model with three components that explained approximately 67% of the original data set for the global model and a conventional PCA model with three components that explained approximately 77% of the original data set were identified.

Loading plots were used to interpret the data and to determine how the variables were interrelated. Fig. 6 shows the loading plots of the global model in the first two reduced dimensions and variable importance in the projection (VIP) plot of a periodic monitoring model. The data in Fig. 6 confirms that the conventional PCA and the periodic monitoring models differentiate the air pollutants variables, which occupy different regions of the plot and exhibit an understandable pattern in which the seven variables are grouped into three clusters.

In the periodic monitoring results, the first cluster was related to temperature and humidity, the second cluster was related to the concentrations of PM_{10} and $PM_{2.5}$, and the last cluster was related to concentrations of NO, NO_X and CO_2. Therefore, if a specific sample was strongly related to any one cluster, then that sample would also have a strong relationship with the pollutants in that corresponding cluster. Clusters could therefore be divided according to the correlation of variables. Since the first cluster was related to PM_{10}, $PM_{2.5}$, NO_2 and CO_2 as is shown in Fig. 6(b), it could have been related to two different variables. The concentrations of $PM_{2.5}$ and PM_{10} were proportional to the quantity of dust that resulted from the inflow of outdoor air and to the movement of passengers. The movement of passengers may have been the primary factor, considering that the Seoul metro is used by about 4.5 million people per day. Carbon dioxide was the component related to the number of real passengers, as it is related to respiration. The second cluster consisted of CO, NO_X and NO, and there was a strong correlation between them. Carbon monoxide and NO_X (specifically nitrogen dioxide) are very similar. It is believed that the effect of the common pollutants can be seen in the process of combustion. The effects of outdoor air, combustion for the operation of the subway and the indoor air used for heating must be taken into consideration as well.

Fig. 6(c) presents a variable importance in the projection (VIP) plot. The variable importance in the projection (VIP) of PCA is defined as follows:

$$\mathrm{VIP} = \sum_{a} (\mathbf{w}_{ak})^2, \tag{4}$$

where $\mathbf{w}_{ak}$ is the weight vector of the MPCA model. The VIP is calculated from the weight vector of the PCA model and the percentage that is explained by the dimension of the model. The VIP can be considered as a measure of how much a certain input corresponds to the samples. Thus, important inputs based on the VIP value can be selected. The VIP is the sum over all model dimensions of the

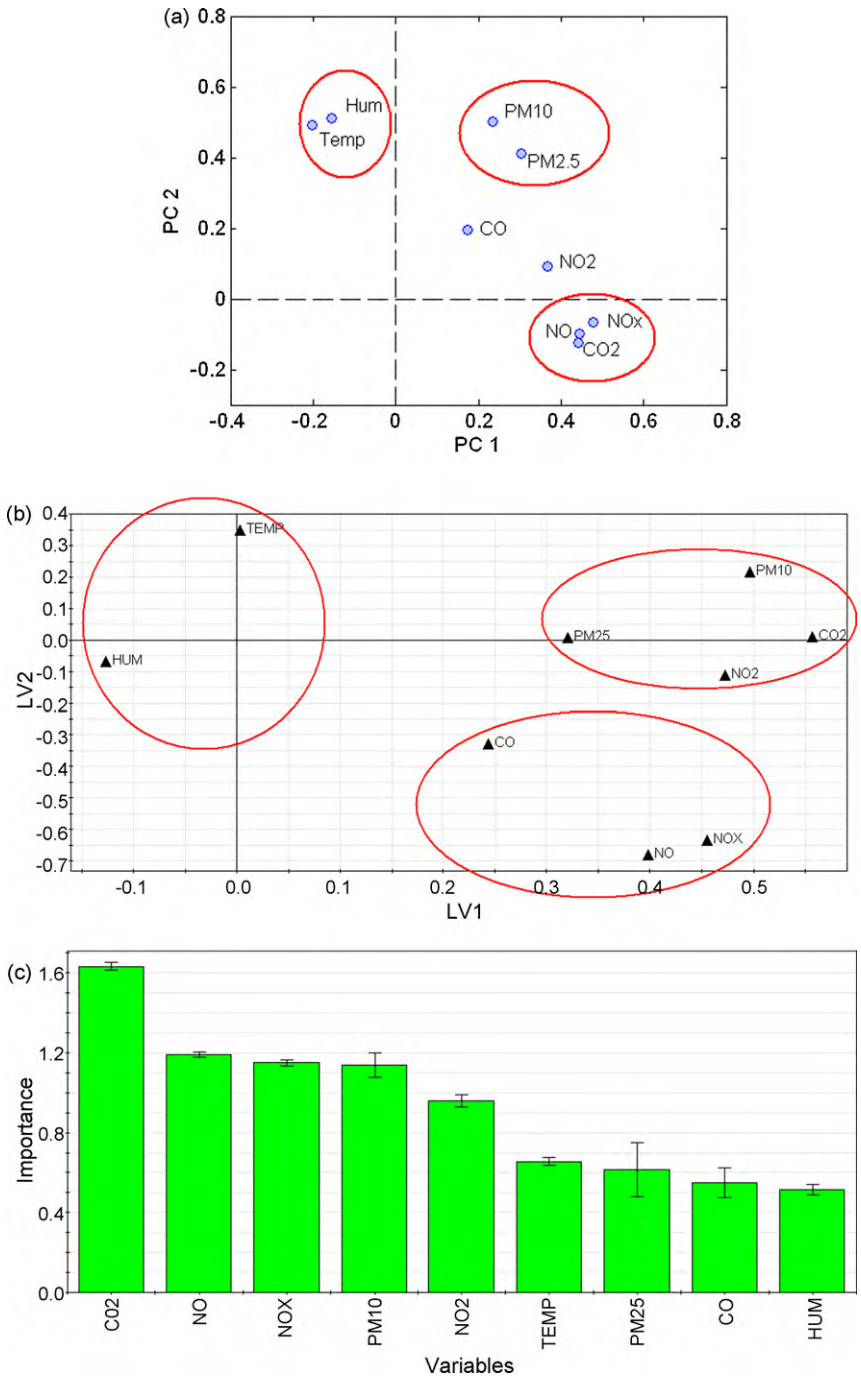

Fig. 6. The loading and VIP plots of the global model for normal condition data during the training period: (a) loading plot of the conventional PCA, (b) loading plot of the periodic monitoring method and (c) VIP plot of the periodic monitoring method.

contributions. Fig. 6(c) is representative of the variables that gradually affect and fit the periodic monitoring model. Therefore, it is confirmed that these variables are effective. Seven air pollutants were determined to be strongly effective variables that belonged to two clusters in the loading plot. More specifically, it was verified that the second cluster was the most effective cluster, and that CO was the most effective variable. The last cluster was related to temperature and humidity. If the temperature is increased in the subway tunnels then the humidity also increases, because the quantity of the vapor from the evaporation of sweat exceeds the quantity of the saturated vapor. Therefore, people in the subway may feel uncomfortable during the summer. Also, humidity and temperature are deeply related to the concentration of formaldehyde. Formaldehyde is present in heat insulators, heating devices, adhesives and smoke [19]. When a person is exposed to formaldehyde, it can stimulate the respiratory system and cause symptoms including a headache, drowsiness, insomnia and asthma, etc.

We monitored conditions for a test period from February to July 2008 in order to investigate the possibility of monitoring using a global model. Fig. 7 shows the monitoring results for the test period using the global model for periodic monitoring.

Fig. 7(a) presents a score plot of indoor air pollutants in the PC_1–PC_2 plane using the global model. It was possible to analyze the indoor air pollution history in the subway station because data with similar pollution conditions tend to cluster together in distinct regions within the reduced space of the global model. In

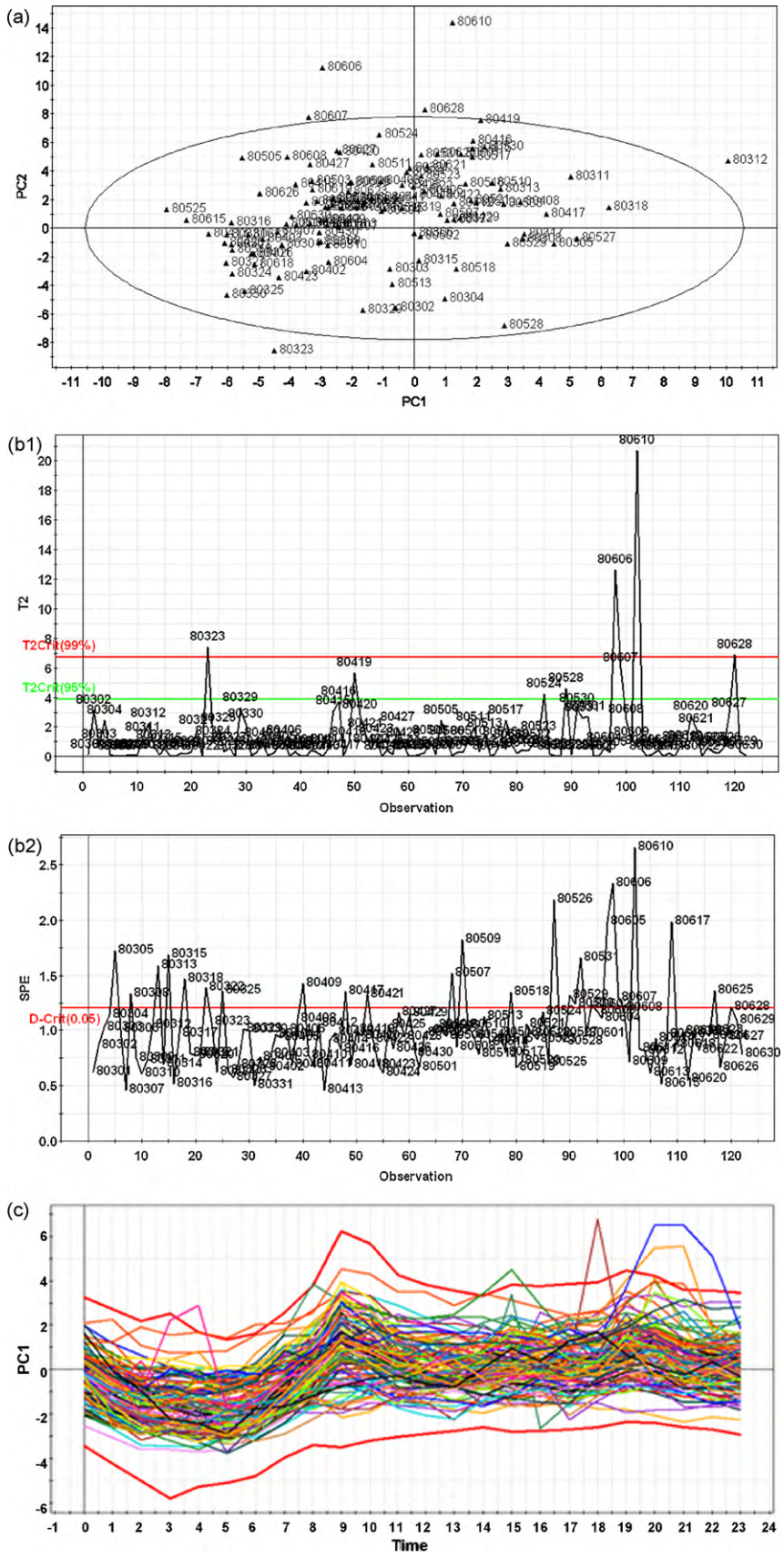

Fig. 7. The monitoring results for the test period using a global model for the periodic monitoring method: (a) score plot, (b) T^2 and Q chart and (c) time plot of the first principal score.

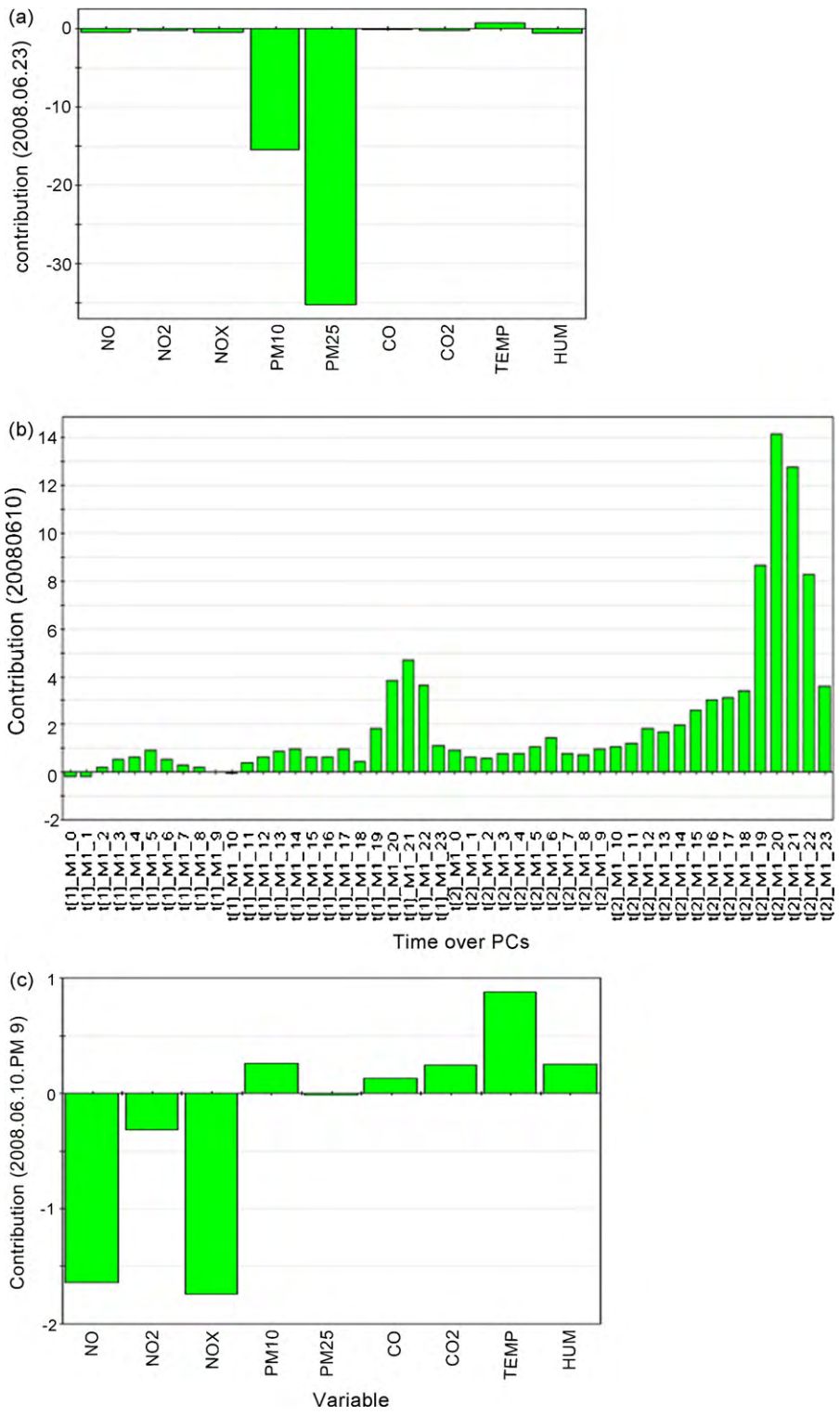

Fig. 8. The contribution and SPE plots for abnormal air quality data obtained from the global model: (a) a contribution plot of the conventional PCA, (b) a time contribution plot for the periodic monitoring model, and (c) a contribution plot for the periodic monitoring model.

addition, it is possible to survey how the points develop as a function of time using SMART charts with the T^2- and Q-statistics, as shown in Fig. 7(b). This overview includes a summary of all of the air quality variables and all of the model dimensions, and it can be used to detect strong deviations in the systematic part of the data. The Hotelling test statistic for multivariate normality, T^2, and the SPE were observed in 95% of the tolerance region. The observed deviations in the Q value can be used to detect changes in contaminated air quality more rapidly than can univariate monitoring.

The greatest difference between the conventional PCA and periodic monitoring models is that the periodic monitoring model can take into account the periodic characteristics of indoor air pollutants when monitoring air quality. Fig. 7(c) presents a time plot of the first principal score. The normal, abnormal or contaminated air quality can be observed over a specific time period in this plot.

Contribution plots and SPE plots are generally used to focus on and investigate the profiles of sample variables. The contribution plots and SPE plots used in this study are presented in Fig. 8.

The contribution plots compare each study sample to the average sample, and the SPE plots show any variables that affected the difference between the model and the sample. The periodic characteristics of the indoor air pollutants can be taken into consideration in these contribution and SPE plots. As shown in Fig. 8(a), in the case of the conventional PCA, it is only possible to inquire which variables most affect the abnormal status of the air pollutants in the multivariate monitoring model at a specific time and date. However, the variable that most affects the abnormal status over a specific time frame for specific data can be identified if the periodic condition is taken into consideration when monitoring the air quality. These results enable more appropriate control of an environmental management system over time and provide information for the design of the system.

3.3. Seasonal model (spring model)

Diurnal and weekly variations in the data over a season are caused by seasonal sources. Therefore, the seasonal impact should be taken into consideration in order to accurately monitor the indoor air quality for the appropriate management of air quality in subway stations.

It is appropriate to use the seasonal models in order to capture the seasonal variations in each pollution region for an indoor air quality monitoring system with seasonal pollution characteristics. If the pollution data corresponding to different seasonal modes exhibits variations due to different meteorological conditions or environmental changes, then each seasonal model can capture the characteristics of its seasonal pollution region better than a global model. However, this increase is at the cost of poor characterization of the other pollution modes.

The data was classified into four groups in order to study the seasonal effects and to establish a seasonal (spring) model to compare to the global model. The procedure for the modeling and monitoring methods was the same as the procedure noted above, and it was also applied to two different methods: conventional PCA and periodic monitoring. March to May 2007 was selected as the training period for modeling the seasonal model. A periodic monitoring model with three PCs that could explain 67.5% of the variation in the data was identified for the seasonal spring model, while the conventional PCA model could explain 77% of the variation in the original data using three PCs. The number of retained principal components and the cumulative percent variance (CPV) for each seasonal model are shown in Table 1.

Fig. 9 presents the loading and VIP plots of the seasonal model for normal air quality condition data. In Fig. 9(a), two clusters were present in each seasonal model. The results of the first cluster were the same for both the conventional PCA and the conventional monitoring model. PM_{10}, $PM_{2.5}$ and NO_2 were strongly correlated with each other in the results of both the conventional PCA and the periodic monitoring model. The strong correlations among these three variables were confirmed by the results of previous studies [10,20–22]. However, the second clusters were different. It was observed in the results of the periodic monitoring model that NO and NO_X were interrelated, while PCA, NO_X, CO_2, CO and NO were correlated with each other in the case of the conventional PCA, as shown in Fig. 9(a).

VIP plot in Fig. 9(c) presents that CO_2 and PM_{10} were the most effective variables and the second cluster was the most effective cluster in the seasonal model of periodic monitoring model. The correlation among the clusters was different in the seasonal model than in the global model. These results indicate that a different relationship may exist between the variables in each of the variable

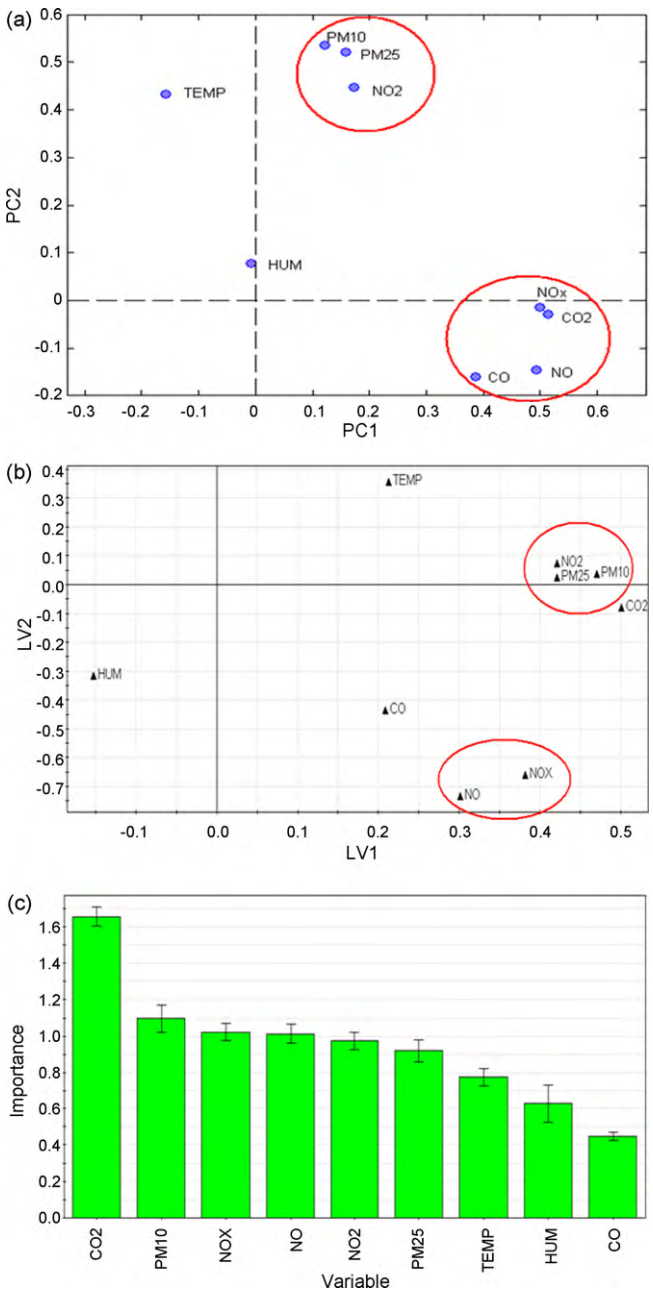

Fig. 9. The loading and VIP plots of the seasonal model for normal air quality data: (a) the loading plot of the conventional PCA model, (b) the loading plot of the periodic monitoring model, and (c) the VIP plot of the periodic monitoring model.

loadings. The PM_{10} and $PM_{2.5}$ concentrations were strongly related in the seasonal model.

The test data set from March to May 2008 was projected onto the seasonal model in order to confirm the monitoring capabilities of the model, and the results are presented in Fig. 10. The status of the air quality can also be observed when monitoring using a normal seasonal model in a 95% tolerance region.

The T^2 and SPE plots shown in Fig. 10, which are similar to the results of the global model, indicate that indoor air quality tends to occasionally become abnormal. In addition, the results indicate that a more persistent trend is establishing itself. The contribution plot of the SPE of abnormal air quality data (not shown in this paper) closely examines the points of difference between each of the samples and models and questions the reasons for the poor air quality using the effective variables of the abnormal air qual-

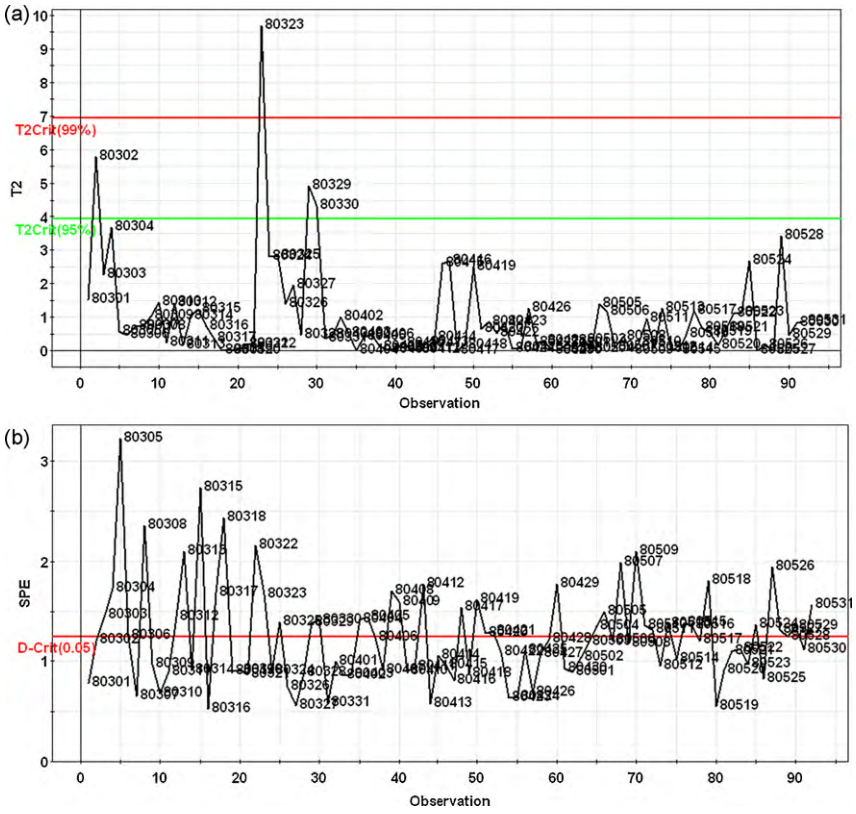

Fig. 10. The T^2 and Q plots for the test period obtained from a seasonal model using a periodic monitoring model.

ity. For example, the air quality on March 23, 2008 (during the spring season), which is noted on both the T^2 and SPE plots, was considered to be abnormal or in a contaminated state, and the concentrations of PM_{10} and $PM_{2.5}$ were particularly high during the evening.

4. Conclusions

Global and seasonal air quality monitoring methods based on multivariate statistical methods were developed and applied to the air pollutant data from a real-time TMS in a subway station in this study. The multivariate air quality monitoring method provided more accurate and reliable results for air pollutants in a subway than did the univariate monitoring method due to the pollutants' multivariate characteristics. More specifically, the seasonal model can detect the abnormal behaviors of air pollutants that lead to unhealthy effects in metro systems that cannot be detected using the global model. In addition, the seasonal models allow us to isolate the characteristics of the seasonal variations for the monitoring of specific air pollutants. The monitoring performance was improved in this study by developing a multiway method to predict the periodic patterns of indoor air pollutants. Better results can be expected by explicitly accounting for the periodic patterns of the air pollutants while applying advanced monitoring and control strategies to the air pollution management system.

Acknowledgements

This work is supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2010-0001860) and the Seoul R&BD Program (CS070160).

References

[1] P.W.G. Liu, Simulation of the daily average PM_{10} concentrations at Ta-Liao with Box-Jenkins time series models and multivariate analysis, Atmospheric Environment 43 (2009) 2104–2113.
[2] D. Aguado, C. Rosen, Multivariate statistical monitoring of continuous wastewater treatment plants, Engineering Applications of Artificial Intelligence 21 (2008) 1080–1091.
[3] J. Lau, W.T. Hung, C.S. Cheung, Interpretation of air quality in relation to monitoring station's surroundings, Atmospheric Environment 43 (2009) 769–777.
[4] B. Lin, B. Recke, J.K.H. Knudsen, S.B. Jorgensen, A systematic approach for soft sensor development, Computers & Chemical Engineering 31 (2007) 419–425.
[5] I. Morlini, Searching for structure in measurements of air pollutant concentration, Environmetrics 18 (2007) 823–840.
[6] J. Ragot, G. Grapin, P. Chatellier, F. Colin, Modeling of a water treatment plant A multi-model representation, Environmetrics 12 (2001) 599–618.
[7] C. Silva, A. Quiroz, Optimization of the atmospheric pollution monitoring network at Santiago de Chile, Atmospheric Environment 37 (2003) 2337–2345.
[8] Y.D. Pan, C.K. Yoo, I. Lee, J.H. Lee, Process monitoring for continuous process with periodic characteristics, Journal of Chemometric 18 (2004) 69–75.
[9] J.C.M. Pires, S.I.V. Sousa, M.C. Pereira, M.C.M. Alvim-Ferraz, F.G. Martins, Management of air quality monitoring using principal component and cluster analysis – part I: SO_2 and PM_{10}, Atmospheric Environment 42 (2008) 1249–1260.
[10] D.M. Markovic, D.A. Markovic, A. Jovanovic, L. Lazic, Z. Mijic, Determination of O_3, NO_2, SO_2, CO and PM_{10} measured in Belgrade urban area, Environmental Monitoring and Assessment 145 (2008) 349–359.
[11] Y.S. Kim, I.W. Kim, J.C. Kim, C.K. Yoo, Multivariate statistical monitoring and local interpretation of indoor air quality in a subway station, Environmental Engineering Science 27 (2010) 901–911.
[12] P. Nomikos, J.F. Macgregor, Monitoring batch processes using multiway principal component analysis, AIChE Journal 40 (1994) 1361–1375.
[13] P. Nomikos, J.F. Macgregor, Multivariate SPC charts for monitoring batch processes, Technometrics 37 (1995) 41–59.
[14] C.K. Yoo, D.S. Lee, P.A. Vanrolleghem, Application of multiway ICA for online monitoring of a sequencing batch reactor, Water Research 38 (2004) 1715–1732.

[15] C.K. Yoo, M.H. Kim, S.J. Hwang, Y.M. Jo, J.M. Oh, Online predictive monitoring and prediction model for a periodic process through multiway non-Gaussian modeling, Chinese Journal of Chemical Engineering 16 (2008) 48–51.
[16] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Windig, R.S. Koch, PLS_Toolbox 3.5, Eigenvector Research. Inc., Manson, 2004.
[17] C.K. Yoo, K. Villez, I.B. Lee, C. Rosen, P.A. Vanrolleghem, Multi-model statistical process monitoring and diagnosis of a sequencing batch reactor, Biotechnology and Bioengineering 96 (2006) 687–701.
[18] L. Eriksson, E. Johansson, N.K. Wold, S. Wold, Multi and megavariate data analysis, in: Principle and Applications, Umetrics AB, Umeå, 2006.
[19] W. Kwon, Behavior of formaldehyde concentration by temperature and humidity of indoor and outdoor in underground shopping center and subway(II), Korean Journal of Sanitation 9 (1994) 67–75.
[20] D.U. Park, K.C. Ha, Characteristics of PM_{10}, $PM_{2.5}$, CO_2 and CO monitored in interiors and platforms of subway train in Seoul, Korea, Environment International 34 (2008) 629–634.
[21] J.C.M. Pires, S.I.V.M.C. Pereir, M.C.M. Alvim-Ferraz, F.G. Martins, Management of air quality monitoring using principal component and cluster analysis – Part II: CO, NO_2 and O_3, Atmospheric Environment 42 (2008) 1249–1260.
[22] C. Rosen, Chemometric approach to process monitoring and control with applications to treatment operation, Ph.D. Thesis, Lund University, Sweden, 2001.