Korean J. Chem. Eng., 32(10), 1960-1969 (2015)
DOI: 10.1007/s11814-015-0042-x

pISSN: 0256-1115
eISSN: 1975-7220

Localized indoor air quality monitoring for indoor pollutants' healthy risk assessment using sub-principal component analysis driven model and engineering big data

Honglan Shi, MinJeong Kim, SeungChul Lee, SeHee Pyo, Iman Janghorban Esfahani, and ChangKyoo Yoo†

Department of Environmental Science and Engineering, Kyung Hee University, Yongin-si, Gyeonggi-do 446-701, Korea
(*Received 17 August 2014 • accepted 25 February 2015*)

Abstract−Indoor air quality (IAQ) in subway systems shows periodic dynamics due to the number of passengers, train schedules, and air pollutants accumulated in the system, which are considered as an engineering big data. We developed a new IAQ monitoring model using a sub-principal component analysis (sub-PCA) method to account for the periodic dynamics of the IAQ big data. In addition, the IAQ data in subway systems are different on the weekdays and weekend due to weekly effect, since the patterns of the number of passengers and their access time on the weekdays and weekend are different. Sub-PCA-based local monitoring was developed for separating the weekday and weekend environmental IAQ big data, respectively. The monitoring results for the test data at the Y-subway station clearly showed that the proposed method could analyze an environmental IAQ big data, improve the monitoring efficiency and greatly reduce the false alarm rate of the local on-line monitoring by comparison with the multi-way PCA.

Keywords: Engineering Data, Big Data, Sub-principal Component Analysis (Sub-PCA), Indoor Air Quality (IAQ), Weekly Effect, Periodic Dynamics

INTRODUCTION

Subway systems are often the most convenient form of public transportation and offer a critical service to city dwellers. However, the millions of people in urban areas who take the subway as their daily transportation suffer long-term exposure to suspended particulate matter (PM) in the underground microenvironment [1]. This exposure may result in serious diseases that are associated with respiratory symptoms and lung malfunction, and even death [2]. Previous studies on the indoor air quality (IAQ) in subway systems have confirmed that hazardous indoor air pollutants accumulate in the subway space due to heavy use, overcrowding, and inadequate ventilation systems [3]. High concentrations of PM were reported in several subway systems, such as the London, New York, Stockholm, and Berlin subway systems [4]. The concentration of fine particulate matter with a diameter less than 2.5 μm ($PM_{2.5}$) in the London subway was 3-10 times higher than with over-ground transportation modes [5]. In Berlin, the concentration of particulate matter with a diameter less than 10 μm (PM_{10}) in the metro was 4.4 times that in a car [6]. Indoor air pollutants directly affect passenger comfort and threaten passenger health [7,8]. Therefore, accurate monitoring and diagnosing of the current IAQ are necessary to ensure the good health of passengers.

In general, IAQ data in subway systems have periodic patterns, such as diurnal and weekly cycle variations, since subway utilization patterns, including the number of passengers and train schedules, are varied throughout a day or a week [9,10]. To consider the hourly variations, the IAQ data in the subway system were formed into a three-dimensional matrix with sample numbers, measured variables, and sample time. Principal component analysis (PCA) is one of the most popular multivariate analysis techniques because it can project data onto a lower dimensionality space to reduce the dimensionality of variables [11]. However, PCA can only be applied in two-dimensional data matrices. Multi-way principal component analysis (MPCA) is an expanded method of PCA that is widely applied to monitor and analyze batch processes with three-dimensional data. Since MPCA regards all of the data as one single object, it cannot take time correlation into account in the IAQ processes [12,13]. Therefore, the monitoring in MPCA would be inefficient with false alarms or missed alarms for the IAQ data from subway systems. To overcome the limitations of the MPCA method, we proposed a new IAQ monitoring method, called sub-principal component analysis (sub-PCA), which is an extension of static PCA to the time slices. Sub-PCA can take the time variations in batch or multi-stage processes into consideration, since sub-PCA can build the statistical limits at each time slice [12]. In addition, weekly models should be developed using the weekday dataset and the weekend dataset to take the weekly effects into account.

The proposed weekly sub-PCA monitoring was applied to the real time IAQ data measured at a subway station in the Seoul metro, Korea, and it had monitoring performance superior to the MPCA.

The theories of the multivariate analysis of variance (MANOVA), MPCA, and sub-PCA are presented in the second section. The third section introduces the motivation of the proposed method, the proposed method, and the materials used in this study. The results and discussion are presented in the fourth section. Finally, the last sec-

†To whom correspondence should be addressed.
E-mail: ckyoo@khu.ac.kr

tion addresses the conclusions of this article.

THEORIES

1. Multivariate Analysis of Variance (MANOVA)

To evaluate the statistical significance of the periodic characteristics in the IAQ process data, the MANOVA test was applied to analyze the weekly characteristics. The MANOVA compared the variance between the two populations to investigate whether the population mean-vectors had changed [14].

A vector of the observations may be described as follows:

$$x_{lj} = \bar{x} + (\bar{x}_l - \bar{x}) + (x_{lj} - \bar{x}_l) \quad (1)$$

where l=1, 2, …, g (g is the number of populations), j=1, 2, …, n (n is the size of samples), and each sample has p variables. $\bar{x}$ is the overall sample mean, $(\bar{x}_l - \bar{x})$ is the estimated treatment effect, and $(x_{lj} - \bar{x}_l)$ is the residual.

If the number of variables in a population is greater than or equal to 1 ($p \geq 1$) and the number of populations is two (g=2), then the value of the F-test statistic can be calculated with the following formula:

$$F = \left(\frac{\sum_{l=1}^{g} n_l - p - 1}{p} \right) \left(\frac{1 - \sqrt{\Lambda^*}}{\Lambda^*} \right) \sim F_{p, \sum_{l=1}^{g} n_l - p - 1}(\alpha) \quad (2)$$

where Λ^* (Wilk's lambda) is calculated as follows:

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|} \quad (3)$$

where $|\mathbf{W}|$ and $|\mathbf{B}+\mathbf{W}|$ are the determinant values of $\mathbf{W}$ and $\mathbf{B}+\mathbf{W}$. The detailed calculations of $\mathbf{B}$, $\mathbf{W}$, and $\mathbf{B}+\mathbf{W}$ are shown in Table 1.

The F-test rejects the null hypothesis H_0: $\tau_1 = \tau_2 = \cdots = \tau_g = \mathbf{0}$, where τ_l is the estimated treatment effect, at level $l = \alpha$, if:

$$F = \left(\frac{\sum_{l=1}^{g} n_l - p - 1}{p} \right) \left(\frac{1 - \sqrt{\Lambda^*}}{\Lambda^*} \right) > F_{p, \sum_{l=1}^{g} n_l - p - 1}(\alpha) \quad (4)$$

where $F_{p, \sum_{l=1}^{g} n_l - p - 1}(\alpha)$ is the upper (100α)th percentile of the F-distribution with p and $\sum_{l=1}^{g} n_l - p - 1$ degree of freedom.

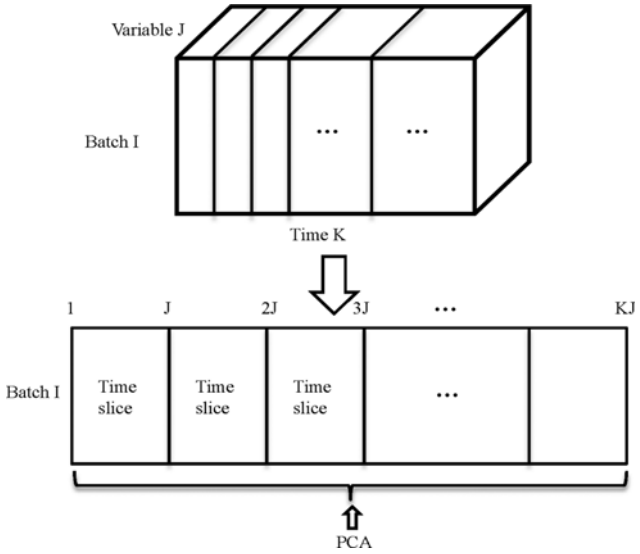

Fig. 1. The conceptual framework of the MPCA method.

2. Multi-way Principal Component Analysis (MPCA)

MPCA can be applied to monitor and analyze batch processes, and the data is usually arranged in a three-dimensional matrix X (I×J×K), where I is the number of batches, J is the number of variables, and K is the number of samples in each batch. The MPCA is expanded by conventional PCA, which is a method to perform conventional PCA on a large two-dimensional matrix unfolded by the three-way data [15]. Fig. 1 shows the conceptual framework of the MPCA method.

Before PCA is performed, the three-dimensional data X (I×J×K) has to be unfolded into the two-dimensional data $\tilde{\mathbf{x}}$ (I×KJ), which is the most meaningful and popular rearrangement for batch analysis and monitoring [16]. After unfolding, PCA can be performed on the two-dimensional data $\tilde{\mathbf{x}}$ (I×KJ) as follows:

$$\mathbf{X} = \sum_{r=1}^{R} \mathbf{t}_r \otimes \mathbf{P}_r + \mathbf{E} = \sum_{r=1}^{R} \mathbf{t}_r \mathbf{p}_r^T + \mathbf{E} = \widehat{\mathbf{X}} + \mathbf{E} \quad (5)$$

where $\otimes$ refers to the Kronecker product, R is the number of retained principal components, $\mathbf{t}_r$ is the relationship among the batches, $\mathbf{P}_r$ is related to both the variable and the sample, and $\mathbf{E}$ is the residual matrix. $\sum_{r=1}^{R} \mathbf{t}_r \otimes \mathbf{P}_r + \mathbf{E}$ represents the three-dimensional decomposition and $\sum_{r=1}^{R} \mathbf{t}_r \mathbf{p}_r^T + \mathbf{E}$ represents the two-dimensional decomposition [15].

Table 1. Calculation of the matrices in the MANOVA

Source of variation	Matrix of sum of square and cross products (SSP)	Degrees of freedom (d.f.)
Treatment	$\mathbf{B} = \sum_{l=1}^{g} \mathbf{n}_l (\bar{\mathbf{x}}_l - \bar{\mathbf{x}})(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})'$	$\mathbf{g-1}$
Residual (error)	$\mathbf{W} = \sum_{l=1}^{g} \sum_{j=1}^{n_l} (\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)(\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)'$	$\sum_{l=1}^{g} \mathbf{n}_l - \mathbf{g}$
Total (corrected for the mean)	$\mathbf{B} + \mathbf{W} = \sum_{l=1}^{g} \sum_{j=1}^{n_l} (\mathbf{x}_{lj} - \bar{\mathbf{x}})(\mathbf{x}_{lj} - \bar{\mathbf{x}})'$	$\sum_{l=1}^{g} \mathbf{n}_l - 1$

For monitoring in multivariable processes, the Hotelling's T^2 statistic and the Q statistic, which is also named the squared prediction error (SPE) statistic, were applied in this study. The T^2 statistic can detect the process faults violating the normal correlation of the variables in the principal component (PC) subspace, and the Q statistic can represent variations due to the remaining PCs, except for the retained PCs. The T^2 statistic and the Q statistic can be represented by the formulas as follows:

$$T_i^2 = t_r^T S^{-1} t_r \sim \frac{R(I-1)}{I-R} F_{R, I-R, \alpha} \tag{6}$$

and

$$Q_i = e_i e_i^T = \sum_{c=1}^{KJ} \mathbf{E}(i, c)^2 \tag{7}$$

where S=diag(λ_1, λ_2, ⋯, λ_l) is a diagonal matrix composed of the eigenvalues of the covariance matrix $\mathbf{X}^T\mathbf{X}$, $F_{R, I-R, \alpha}$ is the *F*-distribution value with R and I−R degree of freedom at the significance level of α, and e_i is the *i*th row of **E**. The distribution of the Q statistic is calculated from the Chi-squared distribution, $SPE_{k,\alpha}=(\nu_k/2m_k)\chi^2_{2m_k^2/\nu_k, \alpha}$, m_k is the mean of the SPE, ν_k is the variance of the SPE, and $\chi^2_{2m_k^2/\nu_k, \alpha}$ is the critical value of the variable χ^2 with a degree of freedom at a significance level α [17].

3. Sub-principal Component Analysis (Sub-PCA)

The data applied in the sub-PCA method is three-dimensional data, $\underline{X}$ (I×J×K), which is the same as in the MPCA method. The conceptual framework of the sub-PCA method is shown in Fig. 2. The three-dimensional process data needs to be unfolded into the two-dimensional data. In Fig. 2, the vertical slice, $\tilde{X}^k$ (I×J), represents a time-slice matrix of the original process data that captures the correlation of the process data at the sampling time k. The static PCA is performed on every time-slice matrix $\tilde{X}^k$ as follows [12]:

$$\tilde{X}^k = T^k (P^k)^T \tag{8}$$

where T^k and P^k are the score matrix and the loading matrix of the time-slice matrix $\tilde{X}^k$ at sampling time k, respectively.

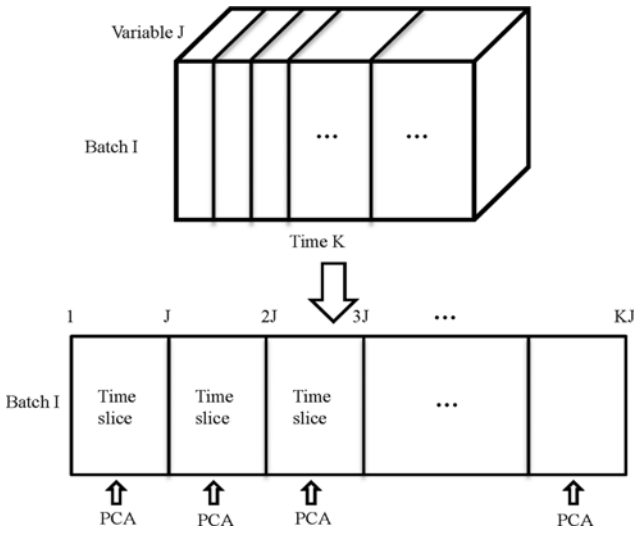

Fig. 2. The conceptual framework of the sub-PCA method (revised from [11]).

To monitor the process data using the sub-PCA method, the loading matrix (P^*) and singular-value diagonal matrix (S^*) are uniformed as follows:

$$P^* = \frac{1}{K}\sum_k \tilde{P}^k \tag{9}$$

and

$$S^* = \frac{1}{K}\sum_k \tilde{S}^k = \mathrm{diag}(\lambda_1^*, \lambda_2^*, \cdots, \lambda_J^*) \tag{10}$$

where $\tilde{S}^k$ is the eigenvalue matrix of the covariance matrix $(\tilde{X}^k)^T \tilde{X}^k$ for the time-slice matrix. The loading matrix P^* can be divided into two parts: the principal component subspace loading matrix ($\bar{P}^*$) and the residual space loading matrix ($\tilde{P}^*$) to reduce the dimensionality of the original data, which is also true for S^*. The optimal number R of retained PCs for each time-slice can be determined by the cumulative percent variance approach. The optimal number of PCs is the first one that can capture a cumulative explained variance of greater than 90% [11]. To monitor the new data (x_{new}) using the sub-PCA method, the T^2 and Q statistics are calculated using the principal component subspace loading matrix ($\bar{P}^*$) and score matrix ($\bar{S}^*$) as follows:

$$T^2_{new} = t^T (\bar{S}^*)^{-1} t \sim [R(I-1)/I(I-R-1)] F_{R, I-1, \alpha} \tag{11}$$

and

$$Q_{new} = e^T e \tag{12}$$

where $t = x_{new}\bar{P}^*$ and $e = x_{new} - x_{new}\bar{P}^*(\bar{P}^*)^T$ [12].

METHODS AND MATERIALS

1. Motivation

The conventional monitoring methods, such as PCA and MPCA, cannot explain the period pattern of IAQ data (i.e., dynamics), since they assume that the periodic pattern of data can be negligible and develop PCA models for the entire data as a single object. Therefore, to account for the changes in the IAQ data correlation from hour to hour (i.e., diurnal dynamics), the sub-PCA method was used for monitoring the IAQ in this study, which divides the data into time-slice matrices and develops PCA in each time-slice matrix.

As mentioned earlier, the IAQ process in a subway system has weekly period patterns. In addition, the difference in the IAQ patterns on the weekdays and weekend is distinct in Fig. 3, which was revealed in the research of Kang et al. [5]. For more convincing evidence, MANOVA was applied in the IAQ process dataset. At a 95% level, the *F* value was 7.52×10^4, which was much larger than the F_{lim} 1.89. This result indicates that the null hypothesis was rejected and there was a significant difference between the weekday data and the weekend data.

Therefore, sub-PCA models for the weekday and weekend IAQ data need to be developed for accurate IAQ monitoring in a subway station. The present study proposes sub-PCA based IAQ monitoring models for the weekday and weekend data.

2. Proposed Weekly Sub-principal Component Analysis (Sub-PCA)

Fig. 4 shows the procedure for the proposed monitoring method

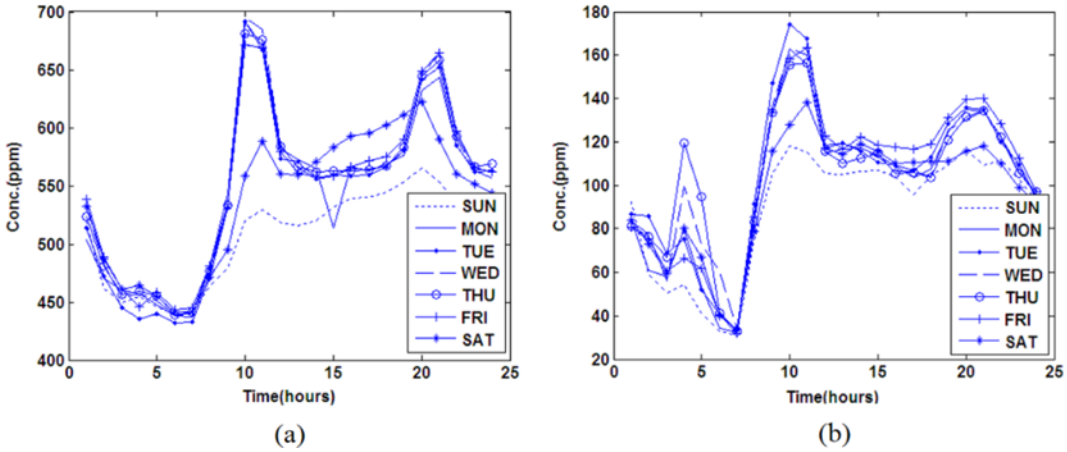

Fig. 3. The periodic change patterns of the indoor air pollutants depending on the variations in the time and day: (a) CO_2 and (b) PM_{10} [12].

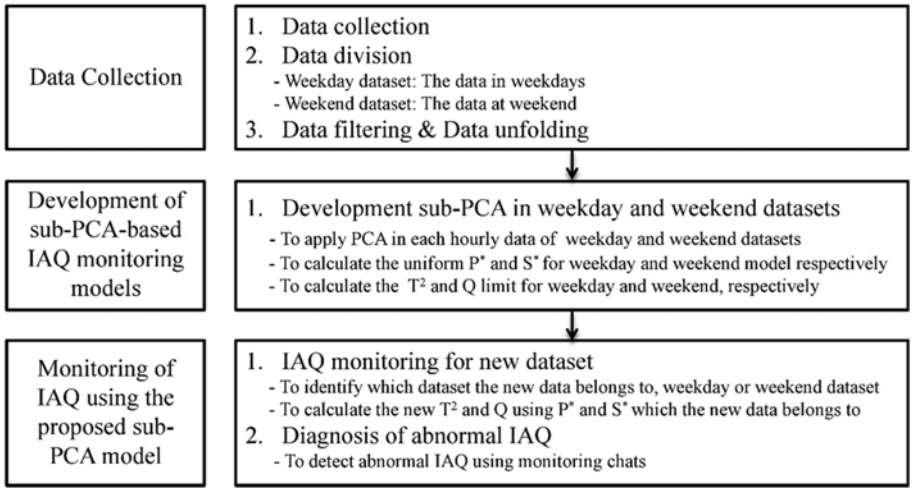

Fig. 4. The procedure for the sub-PCA monitoring method.

of the IAQ in a subway station. This method consists of three parts: (1) data collection, (2) development of the sub-PCA-based IAQ monitoring models, and (3) monitoring of the IAQ using the proposed sub-PCA model.

The data matrix was formed as a three-dimensional matrix after data collection. To construct weekly models, the data were divided into a weekday dataset and a weekend dataset; these two data sets were used as training data. Then, after the outliers were filtered, the three-dimensional datasets (the training data) for the weekdays and weekend were unfolded into the time-slice matrices shown in Fig. 2, where the time-slice matrix represents the variations in the IAQ data within one hour (24 time-slice matrices for 24 hours). The sub-PCA monitoring models were developed by applying the static PCA method to every time-slice matrix. Then, the T^2 and Q statistics of the IAQ data were calculated by using the sub-PCA monitoring models to monitor whether or not the current IAQ was under the normal conditions. To test the superiority of the proposed weekly sub-PCA monitoring method over the conventional methods, the IAQ monitoring results obtained by the weekly sub-PCA were compared with those obtained by the global sub-PCA model and those obtained using the following four different MPCA methods: the weekly online MPCA, the global online MPCA, the weekly offline MPCA, and the global offline MPCA methods.

In this paper, the global model uses the weekday and weekend datasets as a dataset for training, but the weekly model uses the weekday dataset and the weekend dataset for training separately; and the offline MPCA model needs a complete database of batch process for monitoring, while online MPCA model can use an incomplete test data to monitor batch processes in real time [16].

3. Data Collection

The data was collected from April to June, 2008, using a tele-monitoring system (TMS) at the Y-station in Seoul, Korea. There were a total of nine variables, which were $PM_{2.5}$, PM_{10}, nitrogen monoxide (NO), NO_2, nitrogen oxides (NO_x), CO_2, CO, humidity, and temperature which are tele-monitored and used in this study with engineering big data. More details of the IAQ process data are shown in Table 2. As represented in Table 3, the concentrations of PM_{10} and $PM_{2.5}$ were measured by the beta-ray attenuation principle with a corresponding size distribution filter. The concentrations of CO and CO_2 were measured by the non-dispersive infrared radiation absorption of CO and CO_2 molecules at specific wavelengths. The concentrations of NO, NO_2, and NO_x were measured by the

Table 2. Input and output variables of the Seoul subway station data

Variable	Unit	Description	Min	Max	Mean±standard deviation
NO	ppm	Nitrogen monoxide	0.001	0.477	0.0693±0.0753
NO_2	ppm	Nitrogen dioxide	0.001	0.151	0.0513±0.0260
NO_x	ppm	Nitrogen oxides	0.001	0.623	0.1203±0.0974
PM_{10}	μg/m³	Particulate matters with diameters less than 10 μm	3	390	146.3098±56.9984
$PM_{2.5}$	μg/m³	Particulate matters with diameters less than 2.5 μm	2	166	54.0536±25.1318
CO	ppm	Carbon monoxide	0.1	3.4	1.3924±0.7291
CO_2	ppm	Carbon dioxide	400	702	493.1525±49.3398
TEM	°C	Temperature	−3.2	26.4	12.4398±6.1081
HUM	%	Humidity	15.5	84.3	39.5712±14.8354

Table 3. The properties of the measuring instruments

Device (component analyzer)	Detection limit (measuring range)	Measurement accuracy (measurement repeatability)
NO_x analyzer (NA-623)	0.5 ppb (0-1 ppb)	Within ±1% of full scale (FS)
PM_{10} analyzer (SPM-613D)	Less than ±1 μm/m³ (0-0.5/1/2/5 mg/m³)	Less than ±0.5% of FS
$PM_{2.5}$ analyzer (SPM-613)	Less than ±1 μm/m³ (0-0.5/1/2/5 mg/m³)	Less than ±2% of FS
CO_2 analyzer (NDIR gas analyzer)	0.1 ppm (0-5,000 ppm)	Within ±1% of FS

chemiluminescence of the nitro-oxide materials. The detailed properties of the equipment are listed in Table 3. The measurements of each tele-monitoring system (TMS) variable are tele-monitored and used in this study with engineering big data.

The total number of samples was 1776 in this study, among which 1680 samples were used for training (1200 samples on the weekdays, 480 samples at the weekends), and the other 96 samples were used for test. To compare the accuracy of fault detection, four kinds of test datasets (i.e., weekday normal test dataset, weekday abnormal test dataset, weekend normal test dataset, and weekend abnormal

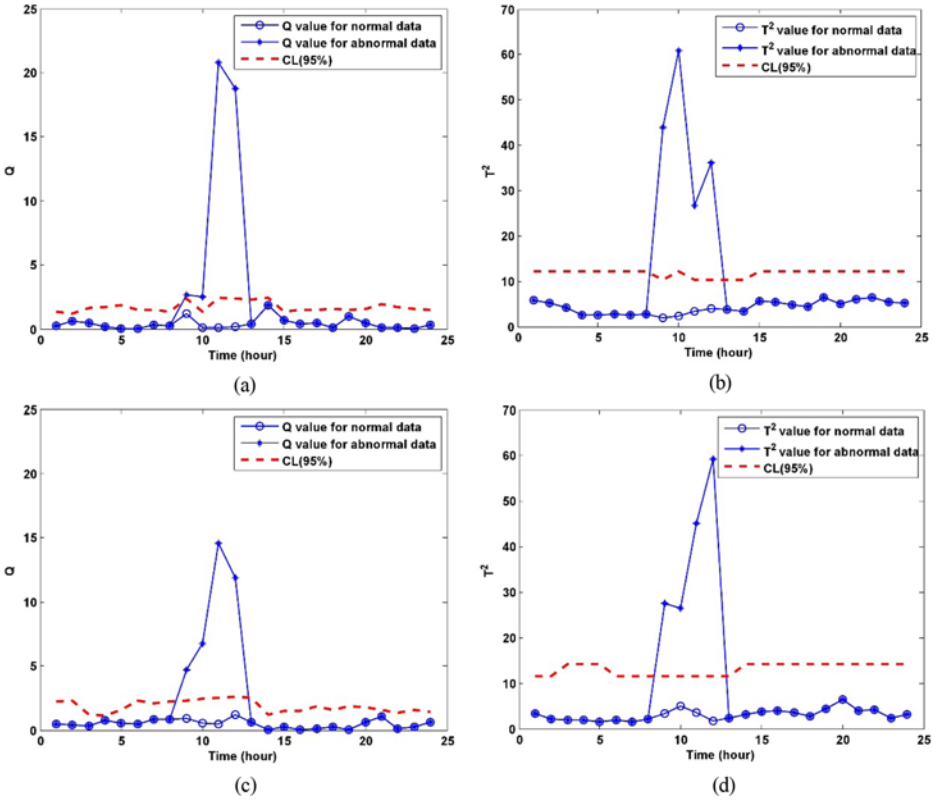

Fig. 5. Monitoring charts for the weekly sub-PCA method.
(a) Q plot for the weekday data, (b) T^2 plot for the weekday data, (c) Q plot for the weekend data, (d) T^2 plot for the weekend data

test dataset) were used. The weekday normal test dataset and the weekend normal test dataset were collected by TMS on a weekday and weekend, respectively. To locate the outliers, bias faults (double average values of the PM_{10} and $PM_{2.5}$ concentrations) were introduced to the samples of the normal test datasets (both weekday and weekend) at 9 a.m., 10 a.m., 11 a.m., and 12 p.m.: subsequently, these datasets with bias faults were used as the weekday abnormal test dataset and the weekend abnormal test dataset.

RESULTS AND DISCUSSION

The monitoring results of the weekly sub-PCA method are presented in Fig. 5. Figs. 5(a) and (b) show the monitoring results of the weekday data in the weekly sub-PCA model. Blue lines with circles in from Fig. 5 to Fig. 10 represent the statistical values for the normal test data, the blue lines with stars represent the statistical values for the abnormal test data, and the red dashed lines are the control limits with 95% confidence. Both the Q plots and the T^2 plots detected the outliers at 9, 10, 11, and 12 for the abnormal data. For the weekend data, the monitoring results are shown in Figs. 5(c) and (d). The Q and T^2 plots detected the outliers at 9, 10, 11, and 12. For the normal data, all of the Q and T^2 plots did not yield any fault alarms. Comparing Fig. 5(a) with (c) or Fig. 5(b) with (d), it can be determined that the control limits for the weekday and weekend data were different. However, in Fig. 6, the control limit of Q (or T^2) for the weekday data is the same as that for the weekend data. In Fig. 6(a), the Q plot cannot detect the outlier at 9, and thus it reveals the difference between the weekly model and the global model. The weekly model had a more exact monitoring capability, because it considered the weekly period characteristics. In contrast, the global model ignored the differences between the weekday data and the weekend data, and it led to the same control limit for the weekday data and the weekend data, and missed the alarm at 9 in time.

Fig. 7 represents the monitoring results of the weekly online MPCA method, and Fig. 8 represents the monitoring results of the global online MPCA method. For the normal data, all of the Q and T^2 plots showed that those data were in normal conditions. However, for the abnormal data, both the weekly and global online MPCA methods had more fault alarms than the sub-PCA method in the Q plots, and some of them were false. This result indicates that when the entire dataset was used to construct the monitoring model without considering the different characteristics of the different time periods, the monitoring models showed more false detections than the sub-PCA models.

The monitoring results of the offline MPCA methods are shown in Figs. 9 and 10. In the offline MPCA, the Q plots and T^2 plots had the same control limit during the 24 hours, instead of using different control limits in different periods like the sub-PCA or online MPCA. These control limits were insufficient for monitoring the abnormal conditions and for diagnosing them in a timely manner. Therefore, the monitoring results of the offline MPCA methods had some false alarms, such as the normal data at 10 in the weekday data in Figs. 9(a) and 10(a), and the abnormal data and

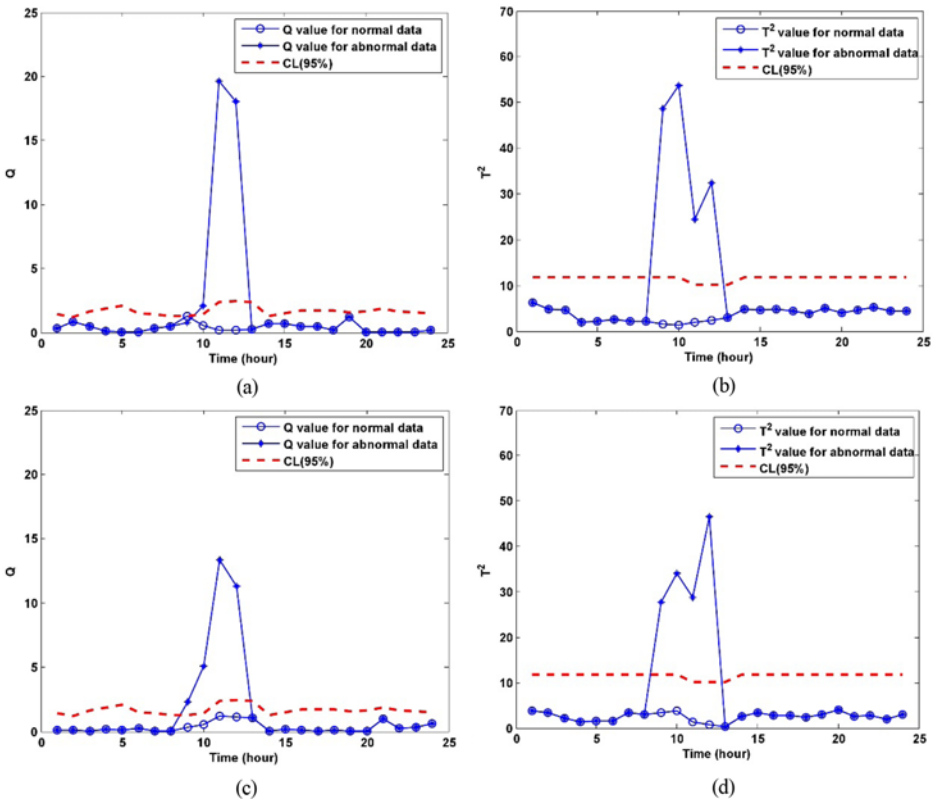

Fig. 6. Monitoring charts for the global sub-PCA method.
(a) Q plot for the weekday data, (b) T^2 plot for the weekday data, (c) Q plot for the weekend data, (d) T^2 plot for the weekend data

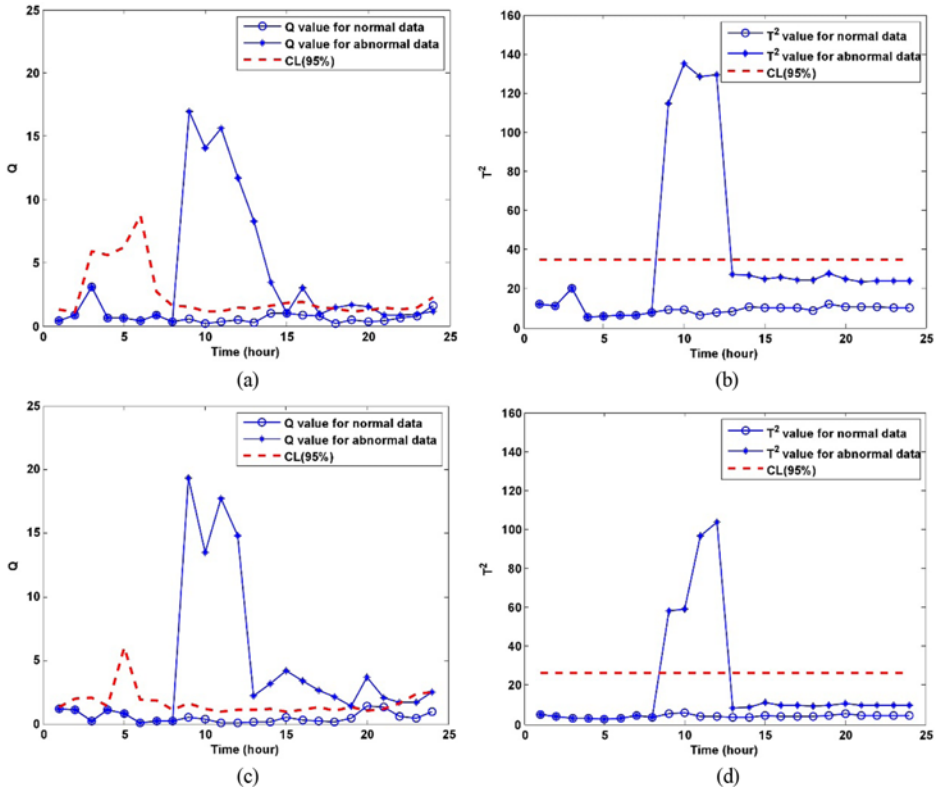

Fig. 7. Monitoring charts for the weekly online MPCA method.
(a) Q plot for the weekday data, (b) T^2 plot for the weekday data, (c) Q plot for the weekend data, (d) T^2 plot for the weekend data

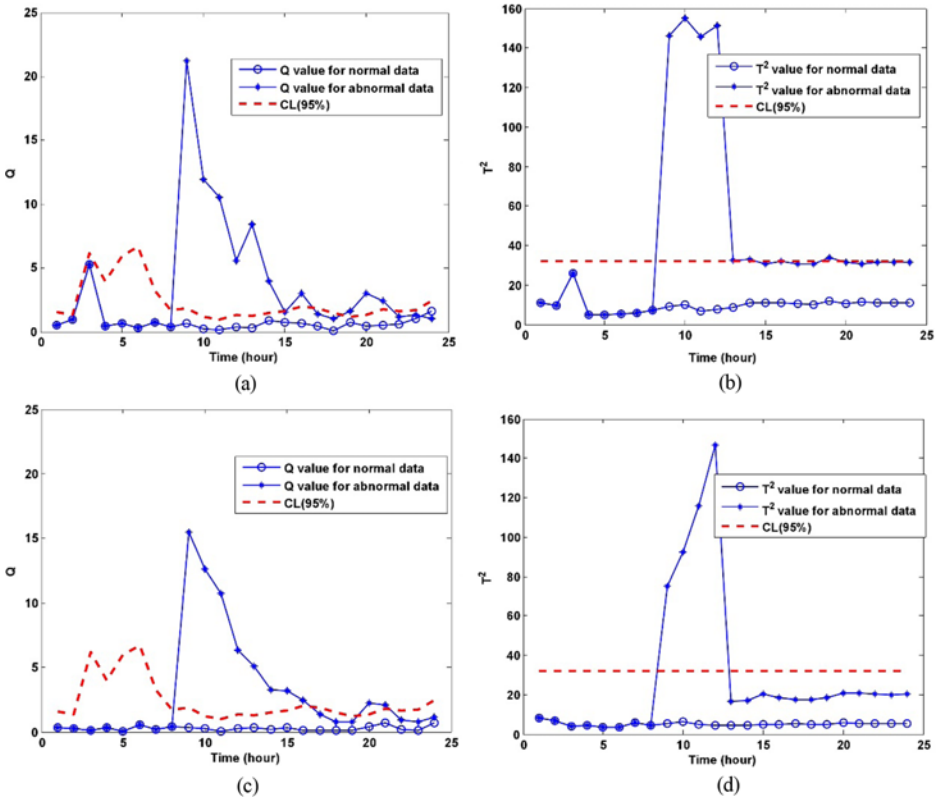

Fig. 8. Monitoring charts for the global online MPCA method.
(a) Q plot for the weekday data, (b) T^2 plot for the weekday data, (c) Q plot for the weekend data, (d) T^2 plot for the weekend data

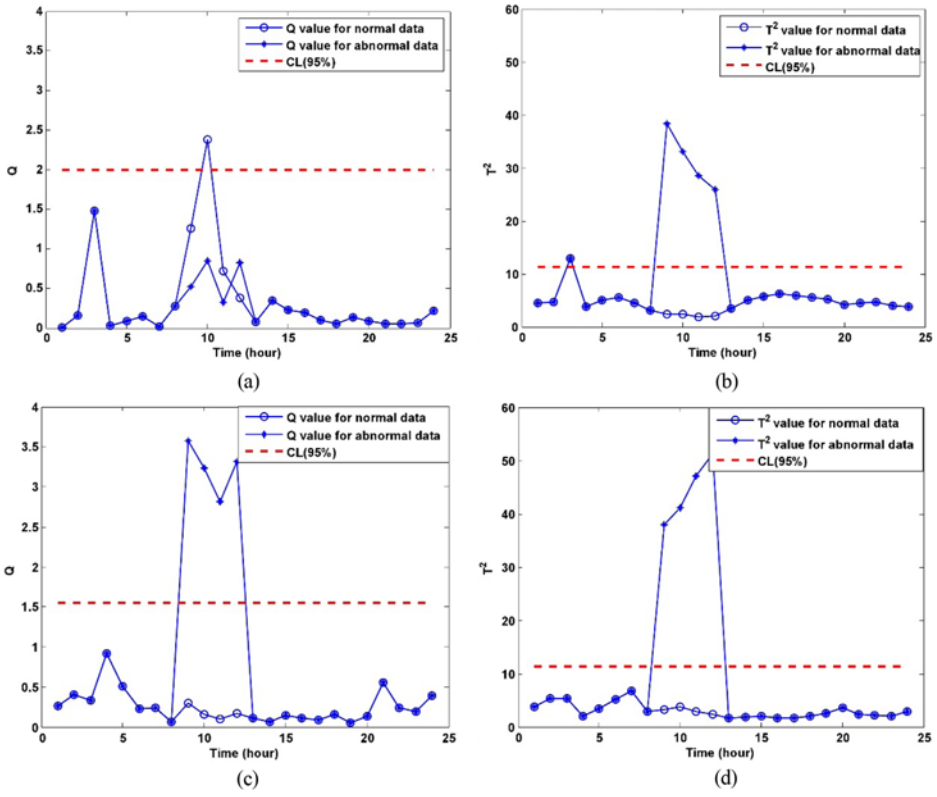

Fig. 9. Monitoring charts for the weekly offline MPCA method.
(a) Q plot for the weekday data, (b) T^2 plot for the weekday data, (c) Q plot for the weekend data, (d) T^2 plot for the weekend data

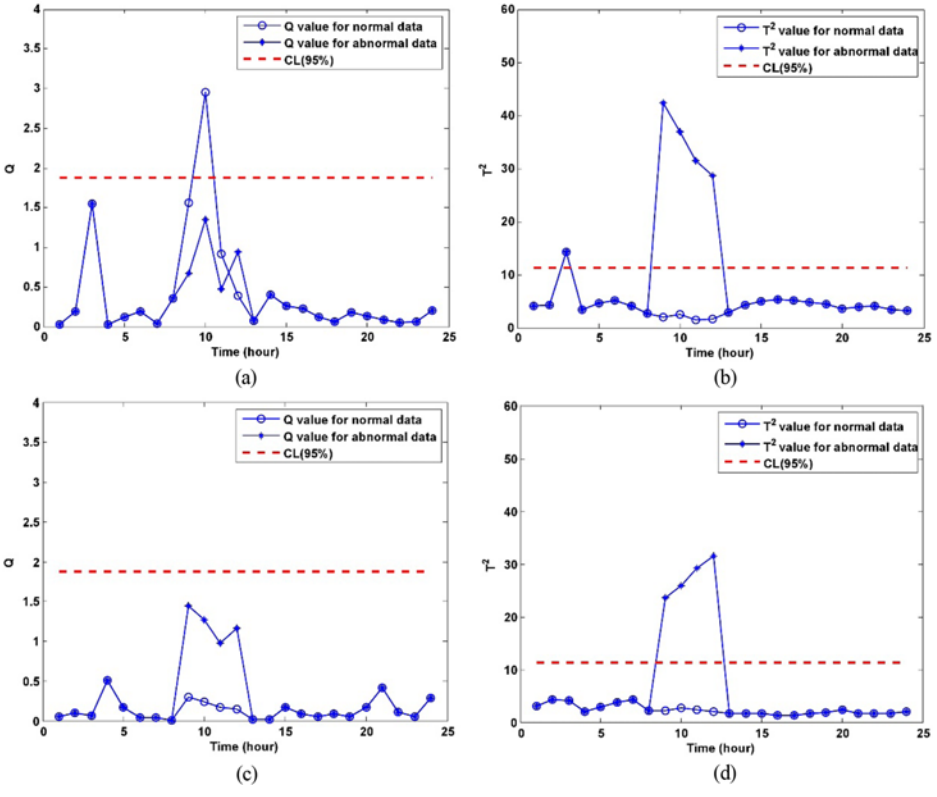

Fig. 10. Monitoring charts for the global offline MPCA method.
(a) Q plot for the weekday data, (b) T^2 plot for the weekday data, (c) Q plot for the weekend data, (d) T^2 plot for the weekend data

Table 4. Comparison of fault detection accuracy of right, false, and missed detection using three monitoring methods

	Weekly sub-PCA		Global sub-PCA		Weekly online MPCA	
	Weekday	Weekend	Weekday	Weekend	Weekday	Weekend
Right detection	4	4	4	4	4	4
False detection	0	0	1	0	6	9
Missed detection	0	0	0	0	0	0
	Global online MPCA		Weekly offline MPCA		Global offline MPCA	
	Weekday	Weekend	Weekday	Weekend	Weekday	Weekend
Right detection	4	4	4	4	4	4
False detection	6	6	3	0	3	0
Missed detection	0	0	4	0	4	4

normal data at 3 in Figs. 9(b) and 10(b). In addition, some of the real fault alarms cannot be detected, such as the outliers at 9, 10, 11, and 12 of the abnormal data in Fig. 9(a) and 10(a) for the weekday data, and 10(c) for the weekend data.

Table 4 lists the fault detection performance of the different monitoring methods, including the numbers of the right detections, false detections, and missed detections. This table shows that the weekly and global online MPCA methods had more false detections than the other methods, while the weekly and global offline MPCA methods had more missed detections than the other methods. The sub-PCA methods had better monitoring capability than the MPCA methods. When comparing the weekly sub-PCA model with the global sub-PCA model, the weekly sub-PCA showed more accurate monitoring results. The weekly sub-PCA method took into consideration the different patterns of the subway system operation between the weekdays and weekend, and the variance between the time periods during a day. Therefore, the weekly sub-PCA method can detect the outliers more accurately than the other methods.

After detecting the fault alarms, the contribution plots could be used to diagnose the variables with the greatest contribution to the abnormal IAQ data. Fig. 11 shows the contribution plot for the abnormal weekday data at 9 in the weekly sub-PCA model. From the figure, it was determined that the $PM_{2.5}$ and PM_{10} had the greatest contribution, and the results were in accordance with those at the beginning of the fourth section. Other fault alarms could also be analyzed through the same approach.

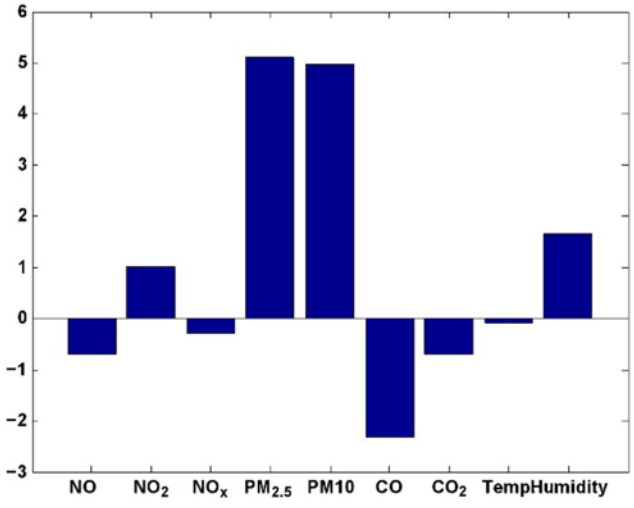

Fig. 11. The contribution plot of the variables for the abnormal air quality data obtained from the weekly sub-PCA model for the weekday dataset.

CONCLUSIONS

IAQ data for indoor air pollutants were localized into weekday and weekend models by using sub-PCA monitoring models. The proposed local based monitoring method was driven to capture the dynamics and periodicity of the IAQ. Since the sub-PCA constructs the PCA model in every time slice rather than treating the entire data as a single object with the MPCA, it is useful to model the dynamics of the IAQ. To tackle the weekly periodicity of the IAQ, we developed weekday and weekend models. The results showed that the sub-PCA monitoring method was able to more precisely detect the abnormal faults of the indoor air pollutants, which are hazardous and high risk to to susceptible or unhealthy groups of people, and was also able to greatly reduce the false alarms in comparison to the MPCA. This result was due to the fact that it could model the time-to-time underlying IAQ variations at different time periods as well as weekly periodicity. As an on-going research work, we have been continuing on researching the environmental big data analysis and the human health risk assessment under varying indoor/outdoor climate conditions.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2015R1A2A2A11001120).

REFERENCES

1. D. S. Grass, J. M. Ross, F. Family, J. Barbour, H. James Simpson, D. Coulibaly, J. Hernandez, Y. Chen, V. Slavkovich, Y. Li, J. Graziano, R. M. Santella, P. Brandt-Rauf and S. N. Chillrud, *Environ. Res.*, **110**, 1 (2010).
2. H. Liu, M. Huang, J. T. Kim and C. Yoo, *Korean J. Chem. Eng.*, **30**,

528 (2013).
3. Y. Kim, J. T. Kim, I. Kim, J. Kim and C. Yoo, *Environ. Eng. Sci.*, **27**, 721 (2010).
4. P. Aarnio, T. Yli-Tuomi, A. Kousa, T. Mäkelä, A. Hirsikko, K. Hämeri, M. Räisänen, R. Hillamo, T. Koskentalo and M. Jantunen, *Atmos. Environ.*, **39**, 5059 (2005).
5. H. S. Adams, M. J. Nieuwenhuijsen, R. N. Colvile, M. A. S. McMullen and P. Khandelwal, *Sci. Total Environ.*, **279**, 29 (2001).
6. M. Nieuwenhuijsen, J. Gomez-Perales and R. Colvile, *Atmos. Environ.*, **41**, 7995 (2007).
7. A. Lai, K. Mui, L. Wong and L. Law, *Energ. and Buildings*, **41**, 930 (2009).
8. R. Kosonen and F. Tan, *Energ. Buildings*, **36**, 981 (2004).
9. O. Kang, H. Liu, M. Kim, J. T. Kim, K. L. Wasewar and C. Yoo, *Indoor Built Environ.*, **22**, 77 (2013).
10. S. Kwon, Y. Cho, D. Park and E. Park, *Indoor Built Environ.*, **17**, 361 (2008).
11. D. Garcıa-Alvarez, *Proceedings of the International Student's Scientific Conference* (2009).
12. N. Lu, F. Gao and F. Wang, *AIChE J.*, **50**, 255 (2004).
13. S. Lee, H. Liu, M. Kim, J. T. Kim and C. Yoo, *Energ. Buildings*, **68, Part A**, 87 (2014).
14. R. A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*, Prentice Hall, Upper Saddle River, New Jersey (2002).
15. Y. Kim, M. Kim, J. Lim, J. T. Kim and C. Yoo, *J. Hazard. Mater.*, **183**, 448 (2010).
16. P. Nomikos and J. F. MacGregor, *AIChE J.*, **40**, 1361 (1994).
17. S. Wang, Y. Chang, Z. Zhao and F. Wang, *Int. J. Control Autom.*, **10**, 1136 (2012).