Korean J. Chem. Eng., **29**(8), 994-1003 (2012)
DOI: 10.1007/s11814-011-0278-z

Analysis and prediction of indoor air pollutants in a subway station using a new key variable selection method

JungJin Lim*, YongSu Kim*, TaeSuk Oh*, MinJung Kim*, OnYu Kang*, Jeong Tai Kim, In-Won Kim***, Jo-Chun Kim****,*****, Jae-Sik Jeon******, and ChangKyoo Yoo*,†**

*Department of Environmental Science and Engineering, Center for Environmental Studies,
**Department of Architectural Engineering, College of Engineering, Kyung Hee University, Gyeonggi-do 446-701, Korea
***Department of Chemical Engineering, ****Department of Environmental Engineering,
*****Department of Advanced Technology Fusion, Konkuk University, Seoul 143-701, Korea
******Seoul Metropolitan Government Research Institute of Public Health and Environment,
Yongmeori2-gil (Juam-dong1), Gwacheon-si, Gyeonggi-do 427-070, Korea
(*Received 11 May 2011 • accepted 26 October 2011*)

Abstract−A new key variable selection and prediction model of IAQ that can select key variables governing indoor air quality (IAQ), such as PM_{10}, CO_2, CO, VOCs and formaldehyde, are suggested in this paper. The essential problem of the prediction model is the question of which of the original variables are the most important for predicting IAQ. The next issue is determining the number of key variables that should be ranked. A new index of discriminant importance in the projection (DIP) of Fisher's linear discriminant (FLD) is suggested for selecting key variables of the prediction models with multiple linear regression (MLR) and partial least squares (PLS), as well as for ranking the importance of input measurement variables on IAQ prediction. The prediction models were applied to a real IAQ dataset from telemonitoring data (TMS) in a metro system. The prediction results of the model using all variables were compared with the results of the model using only key variables of DIP. It shows that the use of our new variable selection method cannot only reduce computational effort, but will also enhance the prediction performances of the models.

Key words: Indoor Air Quality (IAQ), Pollution Prediction, Fisher's Linear Discriminant (FLD), Platform Screen Door (PSD) System, Variable Selection

INTRODUCTION

Subway systems are a major mode of public transportation in many cities around the world. Accordingly, many people living in metropolitan areas spend a considerable amount of time underground in subway environments that subject them to indoor air pollution and directly affect their health [3,5,8,17-19]. To address these concerns, the Korean Ministry of Environment (MOE) drafted an Indoor Air Quality Control in Public Use Facilities Act to control major pollutants, including PM_{10}, CO_2, CO, VOCs and formaldehyde [9]. Advanced monitoring and control strategies for indoor air quality are required to comply with the national law as well as local regulations. Methods to predict indoor air quality (IAQ) are a good way to monitor and control indoor air pollutants since they allow operators to prepare appropriate strategies and ensure safe indoor environments. However, indoor air quality is influenced by a large number of variables, such as the number of passengers, operating conditions, and the concentrations of outdoor air pollutants. Performances of prediction models strongly depend on the variables that are evaluated due to the complicating effects of dimensionality and complexity, as well as collinearity of IAQ. The selection of appropriate IAQ variables is critical to accurately model IAQ; otherwise, predicted IAQ values in subway stations may be different from actual IAQ values. Because there is a big change in IAQ before and after platform screen door (PSD) installation, it is important to carefully select IAQ variables to evaluate IAQ before, during, and after PSD installation.

Previous studies have examined variable selection techniques based on stepwise regression and partial least squares/principal component analysis (PLS/PCA) [10,12,15,20]. Stepwise regression seems to be suitable for selecting variables from a small descriptor pool [12]. PLS and PCA have been used to reduce large descriptor pools to a manageable handful of latent variables related to the actual descriptors via a loading matrix [15]. King and Jackson (1999) studied variable selection in large climate data sets using PCA and suggested four variable selection methods. Ramadan et al. (2001) carried out discriminant partial least squares (DPLS) and genetic algorithms to select variables in environmental soil samples and predicted outcomes using DPLS and artificial neural networks (ANNs). Liu et al. (2003) reported that the best subsets are determined when the cross validation correlation coefficient (q^2) that is predicted during the cross-validation prediction process is used rather than the correlation coefficient (r^2) that is predicted during the modeling estimation process. Kim et al. (2010) suggested that stepwise variable selection should be based on sensitivity analyses, similarity measures, hierarchical clustering, and response surface methods.

In this study, we propose a new variable selection method for models that predict IAQ in a subway station. Our method selects key variables among all original IAQ variables and then predicts

†To whom correspondence should be addressed.
E-mail: ckyoo@khu.ac.kr, ChangKyoo.Yoo@biomath.ugent.be

IAQ based on operating conditions in subway stations. It can reduce the size of the combinatorial problem that results from situations that incorporate a large number of variables. Our new method can simultaneously find a solution to the key variable selection and compute a parsimonious IAQ prediction model while providing the benefits of a reduced computational load.

METHODS

1. Partial Least Squares (PLS)

PLS is a multivariate linear regression algorithm that can handle correlated inputs and limited data [7]. Dimensions of the independent variables (inputs, **X**) and response variables (outputs, **Y**) are reduced by projecting them in directions that maximize the covariance matrix between the input and output variables [22]. The original matrices **X** and **Y** are decomposed as in Eqs. (1) and (2):

$$X = TP^T + E = \sum_{i=1}^{m} t_i p_i^T + E \quad (1)$$

$$Y = UQ^T + F = \sum_{i=1}^{m} u_i q_i^T + F \quad (2)$$

where **p** and **q** are loading vectors that contain information about the relationship between variables, **m** is the number of latent variables, **T** and **U** are the score matrices and **E** and **F** are residuals. As PLS simultaneously projects **X** and **Y** variables onto the same subspace, **T** expresses the relationship between the position of one observation on the **X**-plane and its corresponding position on the **Y**-plane. Once the PLS model has been derived, it is important to grasp its meaning. For this, the scores **t** and **u** are considered. They contain information about the observations and their similarities or dissimilarities in **X** and **Y** space with respect to the given problem and model. The **X** and **Y** weights describe the way in which the variables combine to form **t** and **u**, which in turn express the quantitative relationship between **X** and **Y**. Once the PLS model is established, it can be used for prediction, as PLS seeks a maximum covariance model of the relationship between the X- and Y-spaces [13,16,21].

2. Fisher's Linear Discriminant

There are many feature selection methods, such as PCA and PLS.

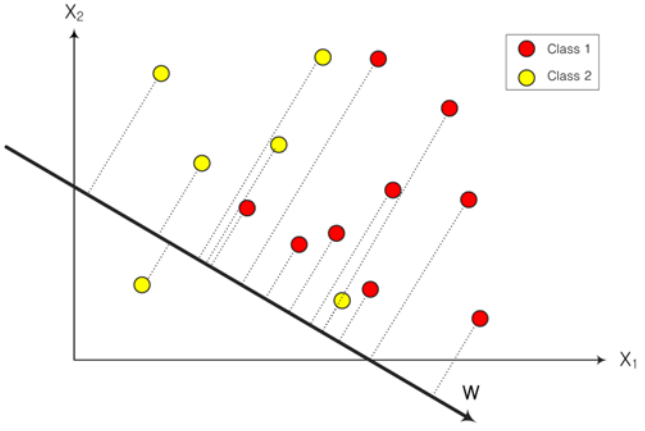


Fig. 1. The concept of optimal projection vector identification by FLD.

PCA identifies components that are useful for representing data, but there is no reason to assume that these components should be useful for discriminating between data in different classes, that is, with different IAQ statuses. FLD analysis is a linear dimensionality reduction technique that maximizes the separation among classes. Where PCA seeks vectors that are efficient for representation, FLD seeks vectors that are efficient for discrimination. Hence, FLD determines a set of projection vectors that simultaneously maximizes the scatter between classes and minimizes the scatter within each class and that maximizes the separability of the data (Fig. 1) [1,2].

Suppose that we have a set of *n*-dimensional samples, $x_1, x_2, \ldots, x_n$, where n_1 and n_2 are the number of samples in class 1 and class 2, respectively. The scatter matrix, S_i, for each class i and the within-class scatter matrix, S_w, are defined as follows:

$$S_i = \sum_{x_i \in X_i} (X_i - \overline{X}_i)(X_i - \overline{X}_i)^T \quad (3)$$

and

$$S_w = \sum_{i=1}^{c} S_i \quad (4)$$

where $\overline{X}_i$ is the mean vector for class i, c is the number of classes and the superscript T denotes the operation of matrix transposition. The between-class scatter matrix, S_B, is defined as follows:

$$S_B = \sum n_i (X_i - \overline{X})(X_i - \overline{X})^T \quad (5)$$

where $\overline{X}$ is the total mean vector. The sum of the between-class scatter matrix and the within-class scatter matrix is equal to the total-scatter matrix, S0, which is defined as follows:

$$S_T = S_B + S_W \quad (6)$$

Under the assumption that S_W is invertible, the projection vector (loading) can be obtained by solving the following optimization problem:

$$\max_{v \neq 0} \frac{v^T S_B v}{v^T S_W v} \quad (7)$$

The FLD vector that maximizes Eq. (7) is equal to the eigenvector w_k of the generalized eigenvector problem

$$S_B w_i = \lambda_i S_W w_i \quad (8)$$

where a generalized eigenvalue of λ_i indicates the degree of overall separability among the classes by projecting the data onto w_k. Because the direction and not the magnitude of w_k is important, the Euclidean norm is usually set at $\|w_i\|=1$ [2,4]. The first FLD vector is the eigenvector associated with the largest eigenvalue, which corresponds to the largest degree of separation among the classes along the direction w_1. With the FLD vectors determined, each sample can then be classified in this reduced FLD space using discriminant analysis [1]. However, it should be noted that FLD implicitly assumes that the population covariance matrices are equal because a pooled estimate of the common covariance matrix is used. If the sample data do not adhere to this assumption, the results of classification will not be satisfactory.

3. Key Variable Selection and Prediction of IAQ

As mentioned above, numerous variables affect IAQ. Collinear-

ity of variables makes it difficult to establish a reliable single model that considers all contributing variables. When the number of variables is large and the fraction of relevant variables is small, prediction models are likely to perform poorly with the large number of variables. The essential problem is the question of which of the original variables are the most important for predicting IAQ. The next issue is determining the number of key variables that should be ranked. A variable important measure, which measures the prediction strength of each variable, should be designed in the variable selection step.

In this paper, discriminant importance in the projection (DIP) of the FLD model which discriminates normal IAQ data from abnormal data, is used as variable important measure. Once the FLD weight vectors are computed, the key variables of the prediction model are selected using Fisher's DIP value which is defined as follows:

$$DIP_k = \sum_i (w_{k,i})^2 \quad (9)$$

where DIP_k is the discriminant importance of the variable (k) in the projection and $w_{k,i}$ are the FLD weights of the key variable that has been used for the IAQ monitoring system. DIP is calculated using Eq. (9) of the weight vector of the FLD model. Thus, important input variables for the prediction model can be selected based on DIP value.

Next, the number of the key variables needs to be decided, since the prediction capability with the selected variables is dependent on how much information one is willing to sacrifice. The eigenvalue-greater-than-one rule of the PCA model is the default option in most statistical packages, and for standardized data the amount of variance extracted by each component should, at a minimum, be equal to the variance of at least one variable. In this paper, the DIP value-greater-than-one rule is used to select the key variable decision. Variables that have DIP values greater than 1 retain only those components that are statistically significant. Thus, the threshold value of $w_{k,i}$ used in this study is 1 (Fig. 5).

Fig. 2 shows the proposed scheme for key variable selection and a prediction model of IAQ. First, IAQ data are measured in subway stations. Any abnormal data that differ from the normal data set, called outliers, are detected and excluded. Multiple linear regression (MLR) and PLS are used to develop the model including all variables. Two multivariate regression models are developed after the essential variables are selected. Overall procedures of the prediction model of IAQ are as follows:

(1) Gather and pre-process the measured data set.

(2) Develop a conventional prediction model using MLR and PLS and all original variables.

(3) Apply the FLD method to the original data and obtain the FLD values.

(4) Calculate the DIP based on the FLD and then select key variables.

(5) Develop MLR and PLS prediction models using the key variables.

(6) Compare the predicted results with results using all original variables and with results using only key variables.

Finally, the recycle loop is used to determine if some of the selected key variables are not appropriate, where a modeling performance of the prediction model with the selected input variables is evaluated and compared. If the prediction evaluation result with the selected variables is not good, the variable selection step should be performed again. The prediction models of MLR and PLS with other

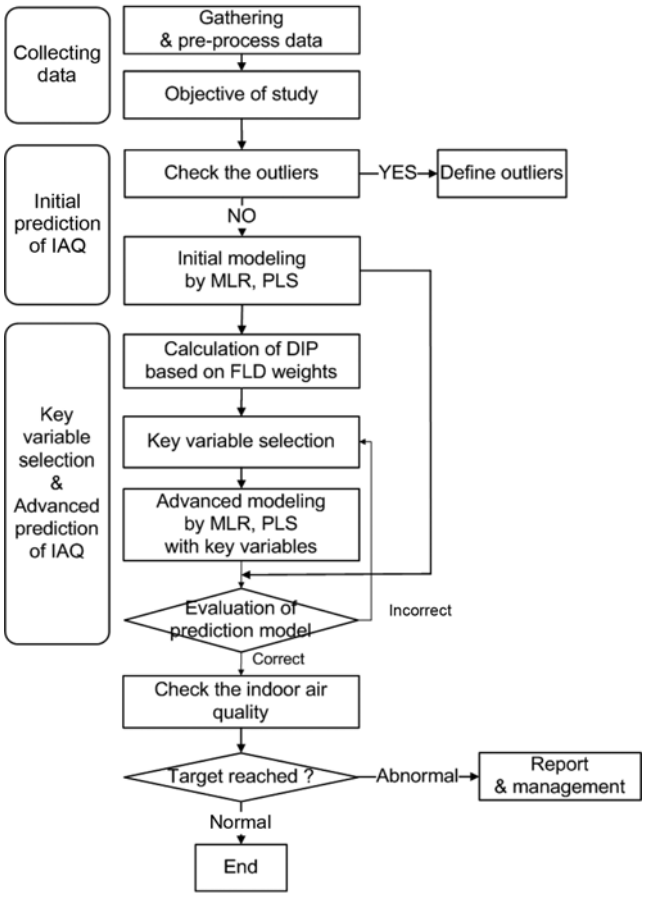


Fig. 2. Proposed scheme for key variable selection and a prediction model of IAQ.

Table 1. Measured indoor air pollutants in the original data

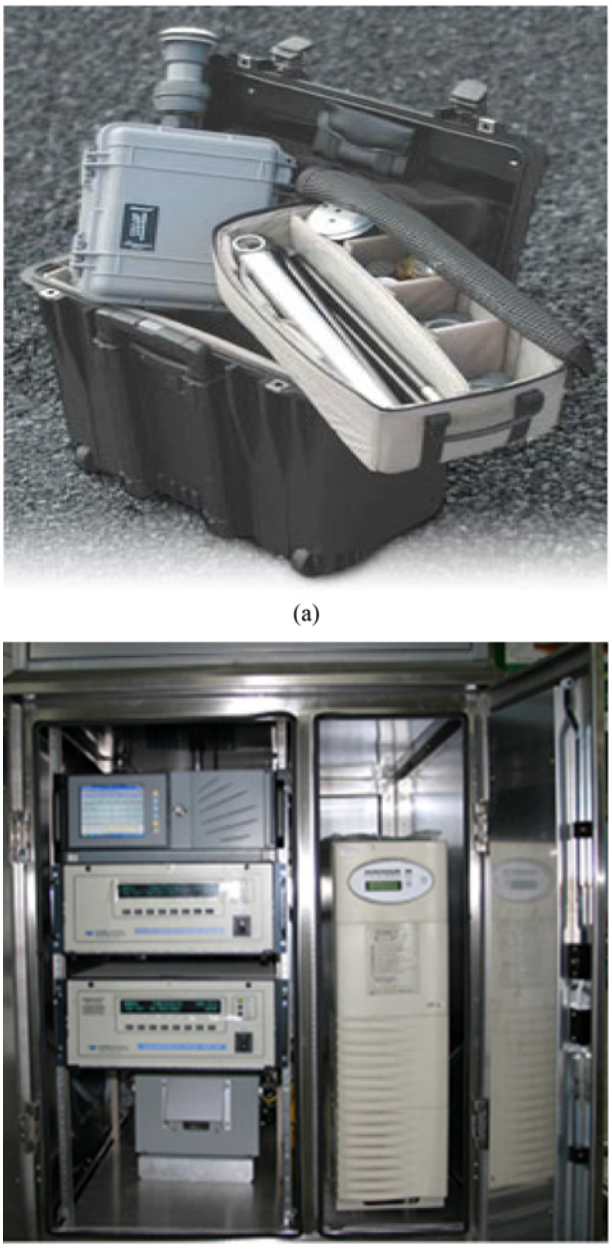
Variable	Mean±SD	Min	Max
Outdoor PM_{10} concentration (μg/m^3)	62.1±35.9	26.2	98.0
Temperature (°C)	8.4±5.7	2.7	14.1
Humidity (%)	56.2±12.4	43.8	68.6
Wind speed (m/sec)	2.1±0.8	1.3	2.9
The concentrations of PM_{10} on the platform (μg/m^3)	115.8±25.4	90.4	141.2
The concentrations of $PM_{2.5}$ on the platform (μg/m^3)	66.2±22.9	43.3	89.1
Number of passengers (persons)	68838.2±10485.3	58352.9	79323.5

input variables are developed. If the evaluation result is appropriate, the key variables of IAQ is measured and monitored (Fig. 2).

4. Data Collection

The proposed method was applied to indoor air data measured off-line at the Y-subway station on line number 3, Seoul Metro, from October 2007 to April 2008. Samples were collected daily and used to create a database with a total of 71 samples. As shown in Table 1, the seven variables for which data were collected were the concentrations of outdoor PM_{10}, temperature, humidity, wind speed, concentrations of PM_{10} and $PM_{2.5}$ on the platform, and number of passengers. Compared to other environmental data sets, Table 1 shows high variability with high standard deviation (SD) and min/max values.

Concentrations of particle matter on the platform were measured once per minute for 20 hours for one day using a mini-volume air sampler (Airmetrics, USA) (Fig. 3(a)). Next, concentrations of outdoor PM_{10} and weather conditions were collected through the telemonitoring system (TMS) of the Seoul Metropolitan Research Institute of Public Health and Environment (Fig. 3(b)) and from internet material from the meteorological office [8]. The number of passengers was determined using Seoul Metro's information system.



(a)

(b)

Fig. 3. (a) Mini volume air sampler (Airmetrics, USA) and (b) TMS system in a subway station.

The dataset was classified into two groups depending on whether a platform screen door (PSD) had been installed to evaluate IAQ before, during, and after PSD installation. With a PSD, subway stations screen the platform from the train. PSDs are full height, total barriers between the station floor and ceiling, while platform edge doors are full height but do not reach the ceiling and thus do not create a total barrier. PSDs can prevent accidental falls off the platform onto the lower track area, improve security and improve indoor air quality control within the station since ventilation and air conditioning are more effective when the station is physically isolated from the tunnel. IAQ concentration showed near normal behavior throughout the entire study period. The first 29 observations describe the period before installation of the PSD system, while the remaining 42 observations describe the period after the PSD was installed.

To determine modeling efficiency, the residual mean square of the error (RMSE) was calculated based on the differences between each real value and its predicted value:

$$\mathrm{RMSE} = \sqrt{\frac{\Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-1}} \tag{10}$$

where Y_i is the real value, $\hat{Y}_i$ is the predicted value, and n is the number of observations [6]. When the RMSE value decreases, the model can be used to accurately predict indoor air quality.

RESULTS AND DISCUSSION

We sought to identify and use key variables for the prediction of IAQ. As mentioned above, several variables affect indoor air quality, but their degrees of influence differ. In addition, it is difficult to implement prediction models that incorporate all of the original variables due to the complexity and dimensionality incurred by using large numbers of variables. Therefore, we performed key variable selection among the original variables using statistical analyses. In the next step, IAQ using only key variables and two regression models, MLR and PLS are predicted. The prediction results using only key variables were compared with those using all of the original variables.

1. Univariate Monitoring of Air Quality

Trends or changes in variables may be monitored or analyzed using a univariate quality index. Fig. 4 shows the univariate quality indexes for seven air pollutants in a subway station. The x-axis represents the number of observations and the y-axis represents the concentrations of indoor air pollutants. Concentrations of outdoor PM_{10} continued to increase even after the 30th observation, indicating that seasonal local sources such as yellow-storm events have a strong influence on particulate matter during this time (Fig. 4).

Concentrations of particulate matter at the platform showed a similar trend up to the 30th observation, after which they began to decrease. Since the PSD system was installed just before the 30th observation, these data suggest that the PSD system affected the

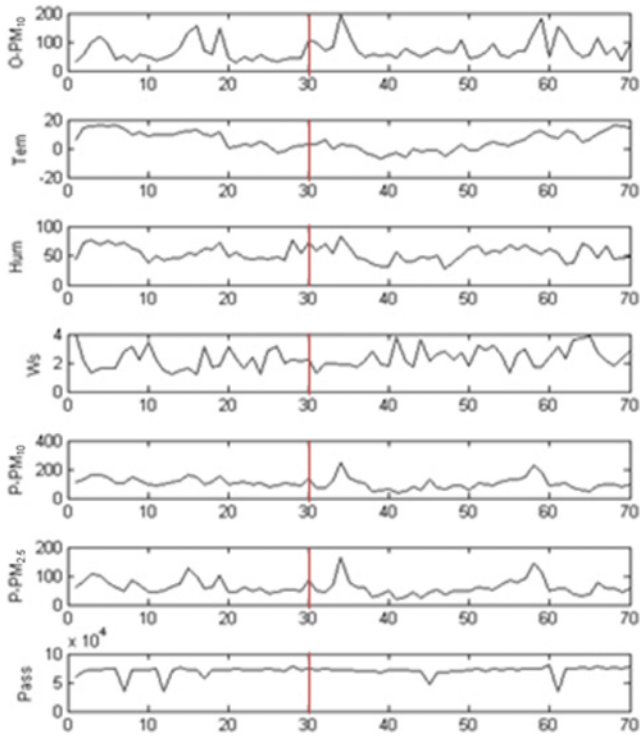


Fig. 4. Univariate quality index of seven indoor variables.

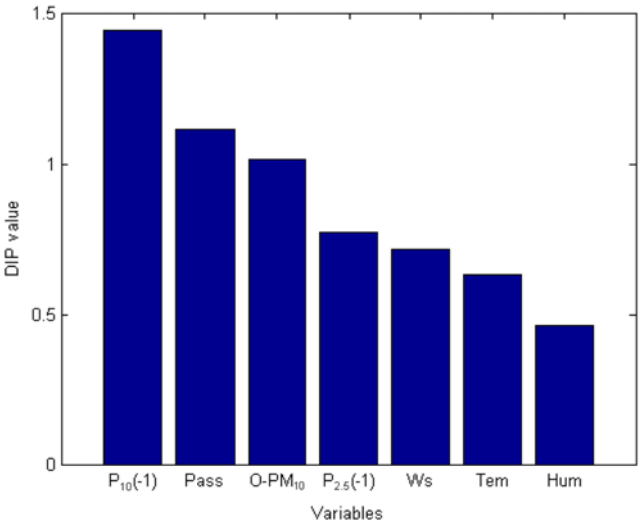


Fig. 5. DIP plot for key variable selection.

* Note) PM_{10}(-1) is the concentration of past PM_{10} at platform, Pass is the number of passengers, O-PM_{10} is the concentration of outdoor PM_{10}, $PM_{2.5}$(-1) is the concentration of past $PM_{2.5}$ at platform, Ws is the wind-speed, Tem is the temperature, and Hum is the humidity.

indoor air quality at the platform by impeding outdoor air pollutants from entering the platform (particulate block effect). Thus, while concentrations of outdoor PM continued to increase, concentrations at the platform decreased. This trend was observed for all of the air pollutants shown in the univariate charts. However, this analysis does not consider correlations among air pollutants. The multivariate method is a good tool to simultaneously monitor IAQ for all pollutants by considering correlations between variables. As mentioned above, the same method can be used for prediction as for monitoring. However, it is necessary to consider only key variables instead of all the original variables.

2. Key Variable Selection

In this step, key variables were selected. Since indoor air pollutants easily accumulate in indoor spaces and then continue to affect IAQ after their accumulations, past the past concentrations of indoor air pollutants must be considered. Therefore, two variables, past concentrations of PM_{10} and current $PM_{2.5}$ at the platform, were added to the original variables. Then, the past concentrations of PM_{10} and $PM_{2.5}$ at the platform, outdoor PM_{10}, temperature, humidity, wind speed, and number of passengers in the subway were selected as input (X) selected variables and the concentrations of PM_{10} and $PM_{2.5}$ at the platform were regarded as output (Y) variables, as listed in Table 2.

Table 2. Descriptions of X and Y variables of IAQ

	Description
X variables	The past concentration of PM_{10} and $PM_{2.5}$ at the platform, outdoor PM_{10} concentration, temperature, humidity, wind speed, the number of passengers in a subway
Y variables	PM_{10} and $PM_{2.5}$ concentrations at the platform

FLD was used to discriminate the original data and obtain FLD weights. The DIP values of X variables were calculated according to Eq. (9). Fig. 5 shows the DIP plot which reflects the importance of each input variable. The important variables based on DIP values are easily identified for the prediction and separation of particulate concentrations measured on the platform. It indicates that the past concentrations of PM_{10} on the platform are the most influential variables. Other major variables, in order of influence, include number of passengers, concentration of outdoor PM_{10}, past concentration of $PM_{2.5}$ on the platform, wind speed, temperature, and humidity. In summary, variables that are related to concentrations of the pollutants, the number of passengers, and outdoor PM_{10} are more important than other meteorological variables for predicting IAQ. It has been known that the number of passengers and concentrations of outdoor air pollutants are known as the main sources that influence indoor air quality [14].

To select the key variables, we used the eigenvalue-greater-than-one rule in this study. The variables with DIP values greater than 1 need to be retained, since the selected variables were statistically significant and defined as the key variables. The past concentrations of PM_{10}, number of passengers, and outdoor PM_{10} were selected as the key variables in this study.

3. A Prediction Model According to the Multivariate Method for Indoor Air Pollutants

When the current IAQ status is determined by the monitoring system, it is possible to manage and control operating conditions to create a safe subway platform environment. However, as mentioned above, prediction of IAQ is also important. When IAQ can be accurately predicted, the operator or engineer is able to manage IAQ by controlling the operating conditions. Therefore, the prediction of IAQ using the selected key variables was the focus of this part of the study.

Outdoor PM_{10}, temperature, humidity, wind speed, the past con-

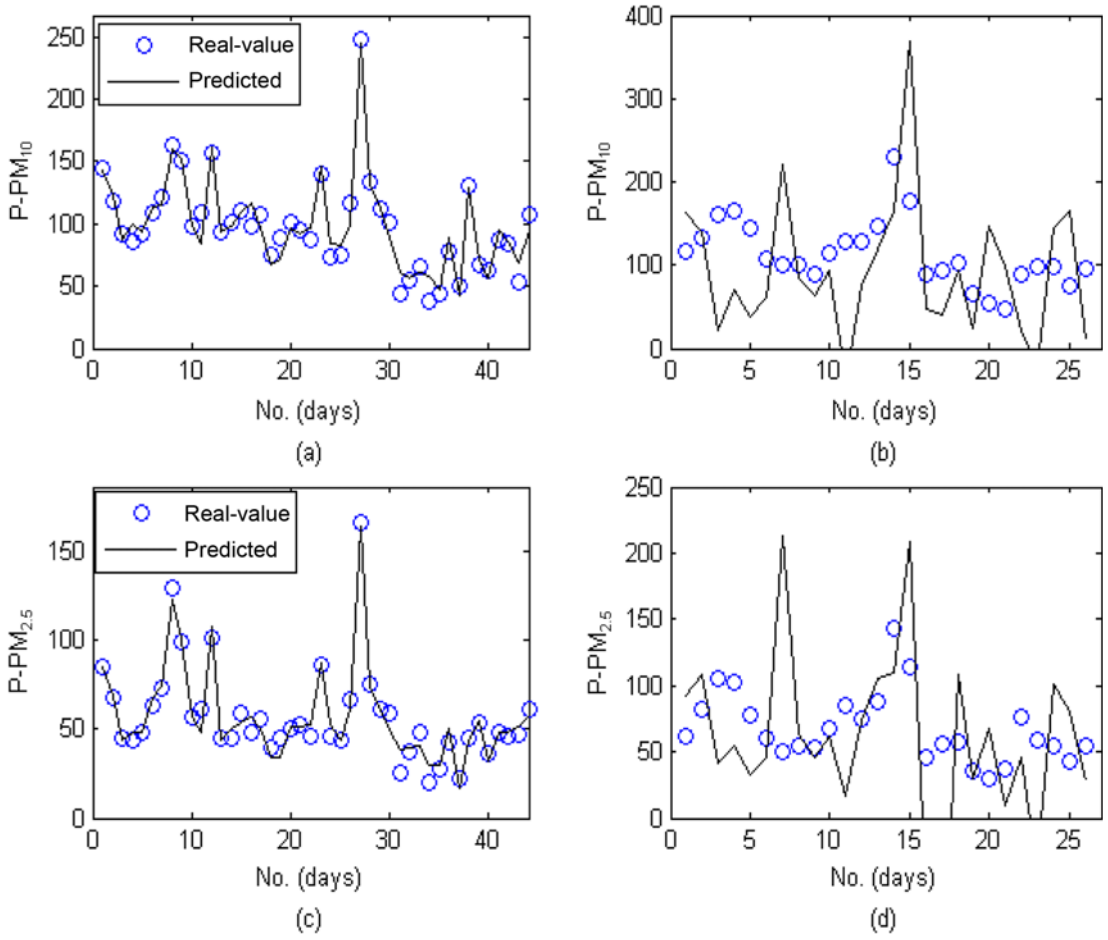


Fig. 6. Prediction results of P-PM_{10} and P-$PM_{2.5}$ using a conventional MLR model with all variables: (a) P-PM_{10} of training data; (b) P-PM_{10} of test data; (c) P-$PM_{2.5}$ of training data and (d) P-$PM_{2.5}$ of test data.

centrations of particulate matter at the platform, and the number of passengers were considered as the X variables and concentrations of PM_{10} and $PM_{2.5}$ at the platform were considered as the Y variables, as shown in Table 2. Forty-four observations were used to establish the model using training data, and the remaining 26 observations were used as test data to examine model efficiency. Both the conventional prediction model including all the original variables and the proposed prediction model including only key variables were used to compare two cases.

3-1. Case 1

For Case 1, the conventional prediction model was performed by MLR and PLS using the original variables. In the MLR model with the quadratic model which considers the square effect as well as interaction effects among the variables, is established. Eqs. (11) and (12) are the quadratic models for PM_{10} and $PM_{2.5}$, respectively.

$$\begin{aligned}\mathbf{PM_{10}}=&-0.1x_1^2-0.05x_2^2+0.1x_3^2-0.5x_4^2-0.5x_5^2-0.5x_6^2-1.4x_7^2-0.1(x_1\times x_2)\\&+0.4(x_1\times x_3)-0.6(x_1\times x_4)+0.2(x_1\times x_5)-0.8(x_1\times x_6)+1.3(x_1\times x_7)\\&-0.3(x_2\times x_3)-0.1(x_2\times x_4)-0.7(x_2\times x_5)+0.6(x_2\times x_6)-0.5(x_2\times x_7)\\&+0.6(x_3\times x_4)+0.4(x_3\times x_5)+0.3(x_3\times x_6)-1.1(x_3\times x_7)-0.8(x_4\times x_5)\\&-0.6(x_4\times x_6)+0.9(x_4\times x_7)-0.6(x_5\times x_6)+1.6(x_5\times x_7)+2(x_6\times x_7)\\&+0.2x_1+0.7x_2-0.2x_3+0.3x_4-0.9x_5+1.1x_6-1.2x_7+0.5\end{aligned}\quad(11)$$

$$\begin{aligned}\mathbf{PM_{2.5}}=&-0.1x_1^2+0.1x_2^2+0.3x_3^2-0.5x_4^2-0.4x_5^2-0.2x_6^2-1.7x_7^2+0.1(x_1\times x_2)\\&-0.03(x_1\times x_3)-0.5(x_1\times x_4)+(x_1\times x_5)-1.6(x_1\times x_6)+2.4(x_1\times x_7)\\&-0.6(x_2\times x_3)-0.3(x_2\times x_4)-0.1(x_2\times x_5)+1.4(x_2\times x_6)-1.4(x_2\times x_7)\\&+0.4(x_3\times x_4)-0.3(x_3\times x_5)+0.2(x_3\times x_6)-0.3(x_3\times x_7)-0.2(x_4\times x_5)\\&+0.1(x_4\times x_6)+0.6(x_4\times x_7)-1.4(x_5\times x_6)+1.7(x_5\times x_7)+1.6(x_6\times x_7)\\&+0.4x_1+0.1x_2+0.4x_3+0.3x_4-1.1x_5+1.5-1.2x_7+0.5\end{aligned}\quad(12)$$

where x_1, x_2, x_3, x_4, x_5, x_6, and x_7 are concentration of outdoor PM_{10}, temperature, humidity, wind speed, number of passengers, past concentration of PM_{10}, and concentration of $PM_{2.5}$ at the platform, respectively. Fig. 6 shows the prediction results using a conventional MLR model with all the original variables. In a time series plot, including the real values and the predicted values. It yielded good results for the training data but poor predictive ability for the test data because the coefficient of each variable was determined for the training data. Thus, this model does not provide a suitable equation.

Another model using the PLS model using three PCs was built. It can explain 67.9% of the X data and 64.8% of the Y data. Fig. 7 shows the results of prediction using a conventional PLS model with all the original variables. RMSE values are presented in Table 3. In the test data, the RMSE values of the MLR model on PM_{10} and $PM_{2.5}$ at the platform were 86.11 and 71.16, respectively. The total RMSE value of the training data was 15.22, and the total RMSE

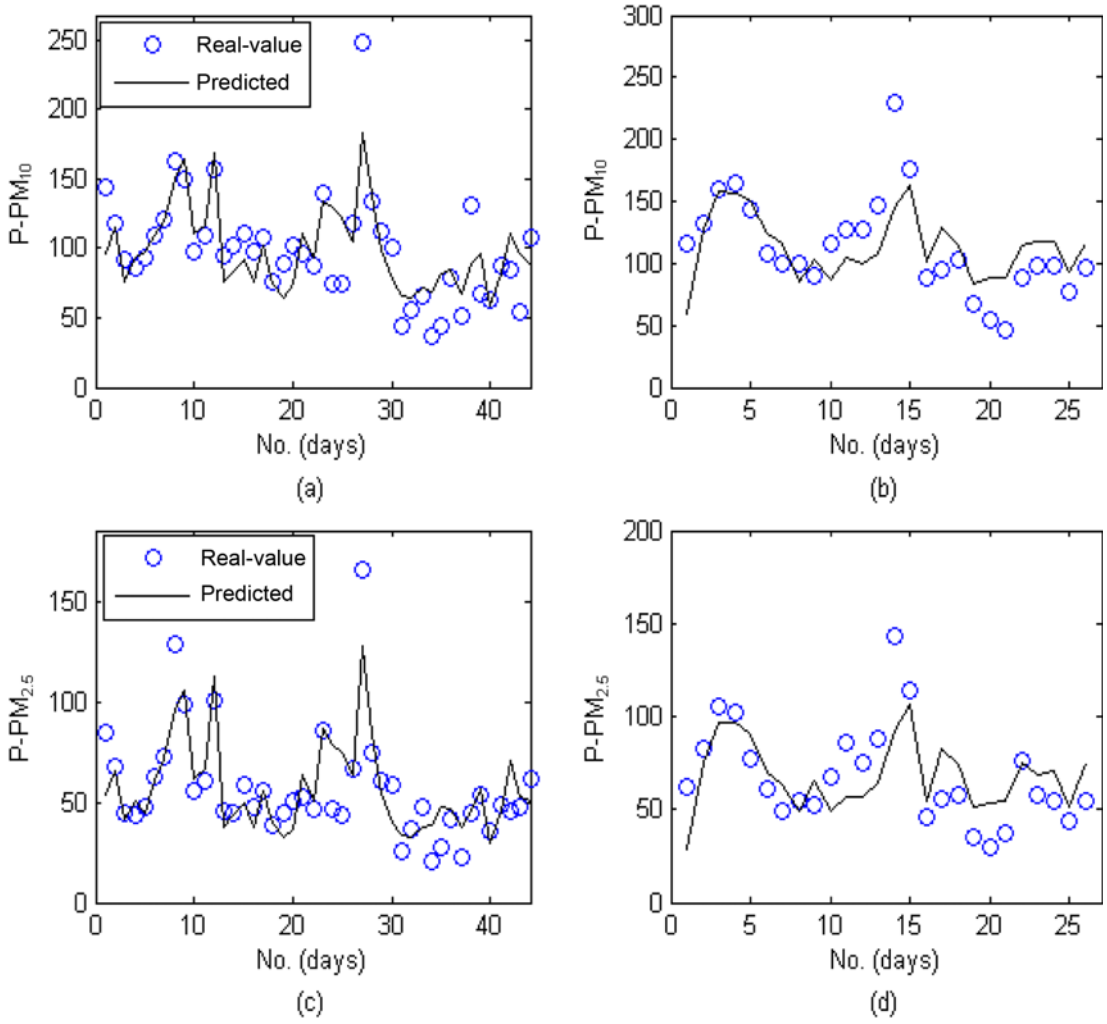


Fig. 7. Prediction results of P-PM_{10} and P-$PM_{2.5}$ using a conventional PLS model with all variables: (a) P-PM_{10} of training data; (b) P-PM_{10} of test data; (c) P-$PM_{2.5}$ of training data and (d) P-$PM_{2.5}$ of test data.

Table 3. RMSE values of MLR and PLS models in prediction results: (a) with all variables and (b) with key variables

	(a) With original variables			
	MLR		PLS	
	P-PM_{10}	P-$PM_{2.5}$	P-PM_{10}	P-$PM_{2.5}$
Training data	9.81	5.41	24.52	14.49
Test data	86.11	71.16	29.80	19.77
	(b) With key variables			
	MLR		PLS	
	P-PM_{10}	P-$PM_{2.5}$	P-PM_{10}	P-$PM_{2.5}$
Training data	22.34	13.89	28.26	17.23
Test data	47.77	21.84	32.06	21.53

Note: P-PM_{10} is the concentration of PM_{10} at platform; P-$PM_{2.5}$ is the concentration of $PM_{2.5}$ at platform

value of the test data was 157.27. In contrast, RMSE values of the PLS model on PM_{10} and $PM_{2.5}$ at the platform were 29.80 and 19.77, respectively. The total RMSE value was 38.97 and the total RMSE value of the test data was 49.57 for the PLS model. These results show that the PLS model is superior to the conventional MLR model for predicting IAQ status.

3-2. Case 2

In a previous step, three key variables with the past concentration of PM_{10} at the platform, the number of passengers, and outdoor PM_{10}, were selected based on their DIP values. For Case 2, two prediction models are performed by MLR and PLS using only selected key variables. The MLR model with the quadratic terms for PM_{10} and $PM_{2.5}$ were built in Eqs. (13) and (14), respectively.

$$\mathbf{PM_{10}}=0.31x_1^2+0.08x_2^2+0.002x_3^2-0.3(x_1\times x_2)-0.23(x_1\times x_3)+0.18(x_2\times x_3)+0.25x_1+0.15x_2+0.28x_3-0.3 \quad (13)$$

$$\mathbf{PM_{2.5}}=0.24x_1^2+0.06x_2^2-0.05x_3^2-0.02(x_1\times x_2)-0.1(x_1\times x_3)-0.02(x_2\times x_3)+0.38x_1+0.2x_2+0.21x_3-0.20 \quad (14)$$

where x_1, x_2, x_3 are the concentration of outdoor PM_{10}, the number of passengers, and the past concentration of PM_{10} at the platform, respectively. Eqs. (13) and (14) yield a much more parsimonious model than Eqs. (11) and (12), with all the original variables. Fig. 8(a) and (b) show the prediction results of the MLR using only key

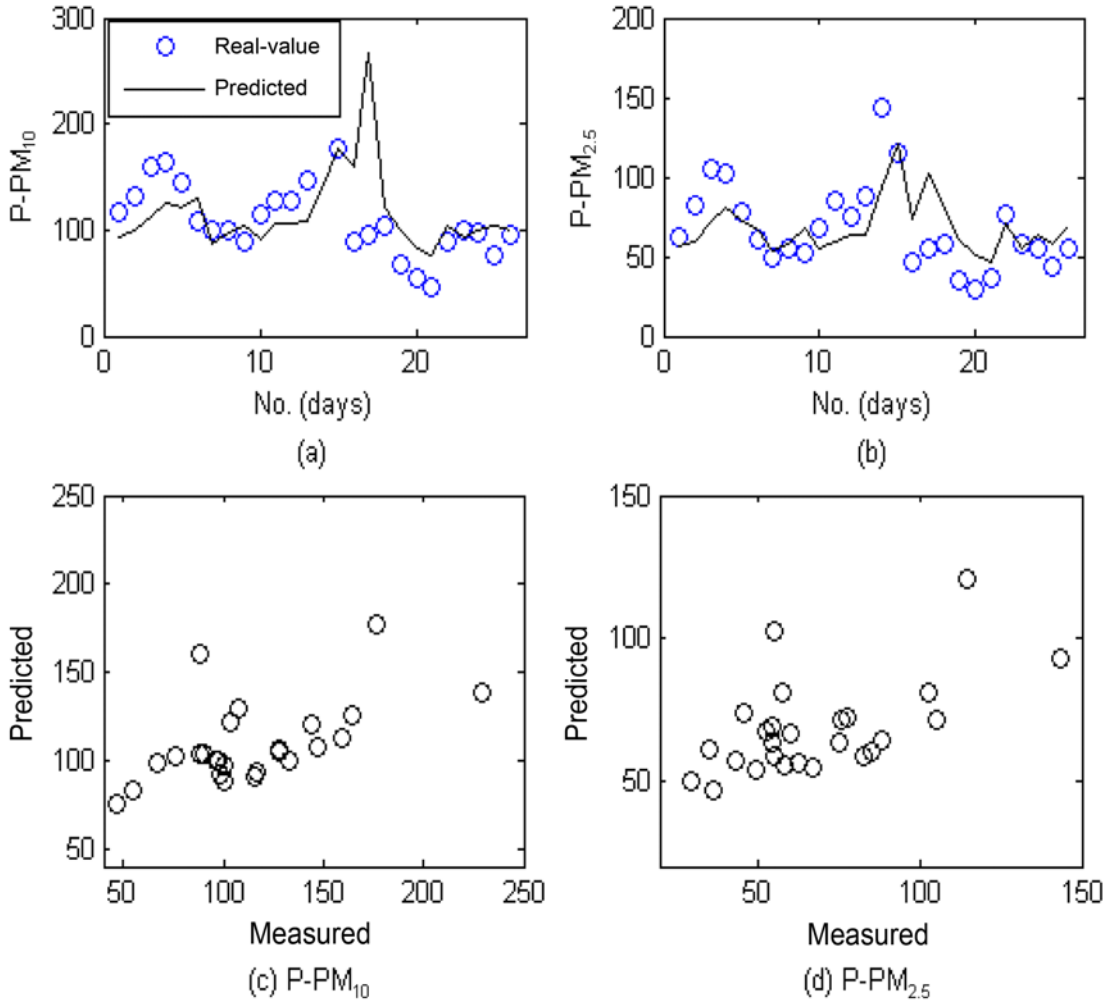


Fig. 8. Prediction results of test data using an MLR model with key variables: (a) time series plot of PM_{10}; (b) time series plot of $PM_{2.5}$; (c) scatter plot of PM_{10} and (d) scatter plot of $PM_{2.5}$.

variables and (c) and (d) are the regression scatter plots. It confirms that the parsimony model can improved the prediction performances.

Fig. 9 shows the results of the prediction results of the PLS model with only key variables. The PLS prediction model can explain 100% of the X data and 52% of the Y data using three PCs. In the test data, the RMSE values of the MLR model on PM_{10} and $PM_{2.5}$ at the platform were 47.78 and 21.84, respectively. The total RMSE value of for the MLR model for the training data was 36.23, and the total RMSE value for the test data was 69.62. In contrast, the RMSE values of the PLS model for PM_{10} and $PM_{2.5}$ at the platform were 32.06 and 21.53, respectively. The total RMSE value was 45.49 and the total RMSE value of the test data was 53.59. It confirms that the PLS model showed better results for predicting IAQ than the MLR model in Case 2 as well as in Case 1.

As shown in Table 3, RMSE values of MLR as well as PLS on the test data set (two case studies) were greatly decreased when selecting the key variables. Compared to the modeling performances with all the original variables, RMSE values of PM_{10} and $PM_{2.5}$ with the MLR model of the selected variables decreased from 86.11 to 44.77 and from 71.16 to 21.84, respectively. On the other hand, the RMSE values of PM_{10} and $PM_{2.5}$ with the PLS model of the selected variables increased from 29.80 to 32.06 and from 19.77 to 21.563, respectively. This indicates that the PLS model has the ability to select the key components of the model by reducing the dimensions of the variables, which results in parsimony of the PLS model structure.

CONCLUSIONS

A new key variable selection method is proposed for the prediction of IAQ in subway stations. The selected key variables can efficiently identify the main pollution sources that affect IAQ in an underground space. The result demonstrates that the RMSEs of the prediction models with only key variables are almost the same as those of the models using all variables, while the PLS prediction model is better for predicting IAQ than the MLR model. The PLS model can identify key variables with a concomitant reduced computational load because it reduces the size of the combinatorial problem resulting from a large number of variables. This study confirms that key variable selection is important for the construction of prediction models of IAQ that includes a lot of input and output variables in a system.

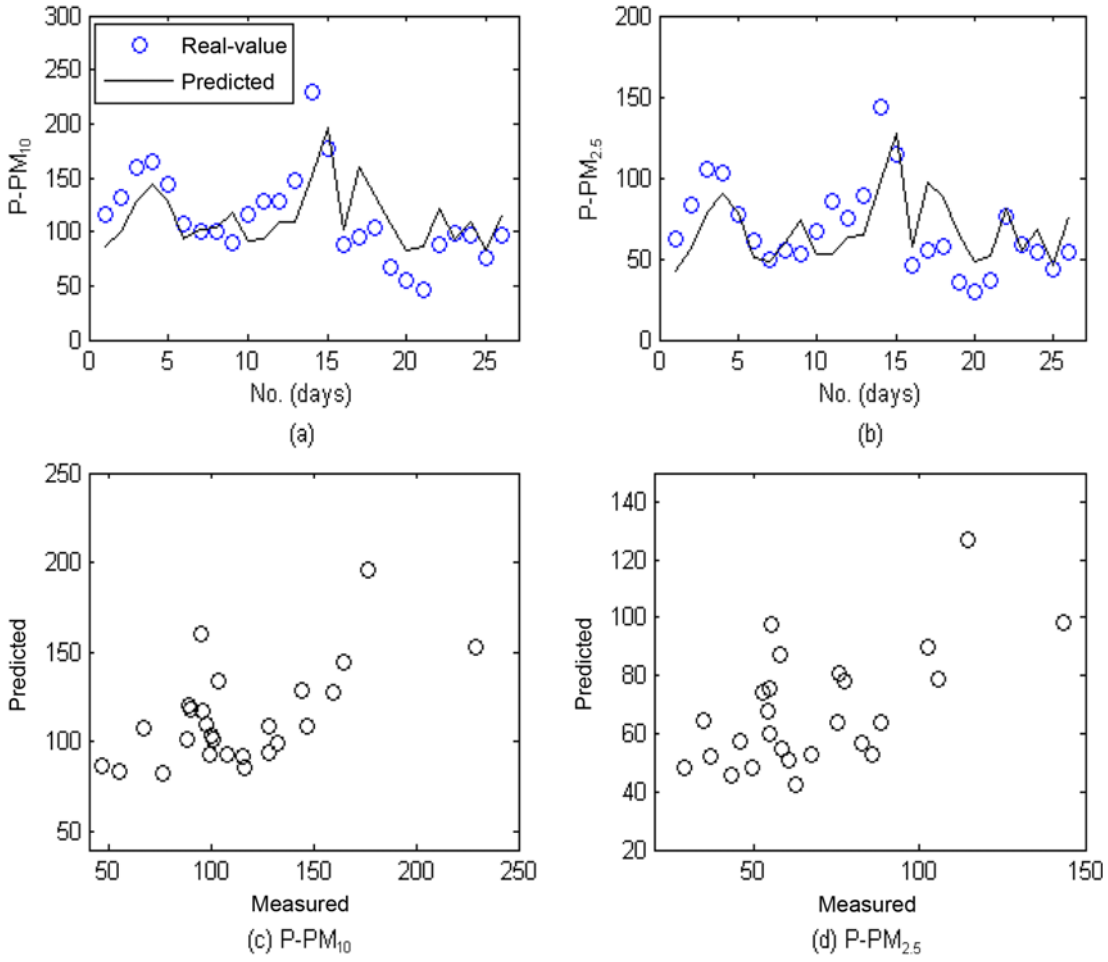


Fig. 9. Prediction results of test data using a PLS model with key variables: (a) time series plot of PM_{10}; (b) time series plot of $PM_{2.5}$; (c) scatter plot of PM_{10} and (d) scatter plot of $PM_{2.5}$.

ACKNOWLEDGEMENTS

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0001031) and Seoul R&BD Program (CS070160).

NOMENCLATURE

E, F : residuals [-]
DIP_k : discriminant importance in the projection (DIP) [-]
m : number of latent variables [-]
p, q : loading vectors that contain information [-]
S_i : scatter matrix [-]
S_W : within-class scatter matrix [-]
S_B : between-class scatter matrix [-]
S_T : total-scatter matrix [-]
T, U : score matrices [-]
w_k : eigenvector [-]
w_{ak} : FLD weights [-]

REFERENCES

1. L. H. Chiang, E. Russell and R. D. Braatz, *Chem. Int. Lab. Syst.*, **50**, 243 (2000).
2. R. O. Duda, P. E. Hart and D. G. Stork, *Pattern classification*, 2nd Ed., John Wiley & Sons, New York (2001).
3. K. Furuya, Y. Kudo, K. Okinaga, M. Yamukki, S. Takahashi, Y. Araki and Y. Hisamatsu, *J. Trace and Microprobe Techniques*, **19**, 469 (2001).
4. T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning*, Springer, USA (2009).
5. S. N. Kang, H. J. Hwang, Y. M. Park, H. K. Kim and C. U. Ro, *Environ. Sci. Technol.*, **42**, 9051 (2008).
6. M. H. Kim, A. S. Rao and C. K. Yoo, *Ind. Eng. Chem. Res.*, **48**, 6363 (2009).
7. M. J. Kim, M. H. Kim, Y. S. Kim and C. K. Yoo, *Multivariate interpretation on the parameter correlation for over-determined system modeling of ASM model*, 10th IWA Conference on Instrumentation Control and Automation (2009).
8. N. J. Kim, S. S. Lee, J. S. Jeon, J. H. Kim and M. Y. Kim, Evaluation of Factors to Affect PM-10 Concentration in Subway Station, *Proceedings of Korean Society for Atmospheric Environment*, 571 (2006).
9. Y. S. Kim, J. T. Kim, I. W. Kim, J. C. Kim and C. K. Yoo, *Environ. Eng. Sci.*, **27**, 721 (2010).
10. Y. S. Kim, M. H. Kim and C. K. Yoo, *J. Hazard. Mater.*, **183**, 441

(2010).
11. Y. S. Kim, M. J. Kim, J. J. Lim, J. T. Kim and C. K. Yoo, *J. Hazard. Mater.*, **183**, 448 (2010).
12. J. R. King and D. A. Jackson, *Environmetrics*, **10**, 67 (1999).
13. T. Kourti, *Anal. Bioanal. Chem.*, **384**, 1043 (2006).
14. H. K. Kwag, K. W. Jin, W. Kim, W. S. Yang, S. J. Choi and D. U. Park; *Kor. J. Env. Hlth.*, **31**, 379 (2005).
15. S. S. Liu, H. L. Liu, C. S. Yin and L. S. Wang, *J. Chem. Inf. Comput. Sci.*, **43**, 964 (2003).
16. J. F. MacGregor, C. Jaeckle, C. Kiparissides and M. Koutoudi, *AIChE J.*, **40**, 826 (1994).
17. L. G. Murruni, V. Solanes, M. Debray, A. J. Kreiner, J. Davidson and M. Davidson, *Atoms. Environ.*, **43**, 4577 (2009).
18. M. J. Nieuwenhuijsen, J. E. Gomez-Perales and R. N. Colvile, *Atmos. Environ.*, **41**, 7995 (2007).
19. A. Paivi, Y. T. Tarja, K. Anu, M. Timo, H. Anne, H. Kaarle, R. I. Mika, H. Risto, K. Tarja and J. Matti, *Atmos. Environ.*, **39**, 5059 (2005).
20. Z. Ramadan, X. H. Song, P. K. Hopke, M. J. Johnson and K. M. Scow, *Anal. Chim. Acta*, **446**, 233 (2001).
21. A. M. Rigol, M. Camps, A. D. Juan, G. Rauret and M. Vidal, *Environ. Sci. Techonol.*, **42**, 4029 (2008).
22. T. Yamamoto, A. Shimameguri, M. Ogawa, I. Hashimoto and V. Kano, Application of Statistical Process Monitoring with External Analysis to an Industrial Monomer Plant, *IFAC Symposium on Advanced Control of Chemical Processes*, 405 (2004).