Korean J. Chem. Eng., **27**(6), 1675-1680 (2010)
DOI: 10.1007/s11814-010-0313-5

Data-driven prediction model of indoor air quality in an underground space

Min Han Kim*, Yong Su Kim*, JungJin Lim*, Jeong Tai Kim, Su Whan Sung***, and ChangKyoo Yoo*,†**

*Department of Environmental Science and Engineering, Center for Environmental Studies,
Kyung Hee University, Seocheon-dong 1, Giheung-gu, Yongin-si, Gyeonggi-do 446-701, Korea
**Department of Architectural Engineering, Kyung Hee University,
Seocheon-dong 1, Giheung-gu, Yongin-si, Gyeonggi-do 446-701, Korea
***Department of Chemical Engineering, KyungPook National University,
1370 Sankyuk-dong, Buk-gu, Daegu 702-701, Korea

(*Received 3 February 2010 • accepted 24 April 2010*)

Abstract−Several data-driven prediction methods based on multiple linear regression (MLR), neural network (NN), and recurrent neural network (RNN) for the indoor air quality in a subway station are developed and compared. The RNN model can predict the air pollutant concentrations at a platform of a subway station by adding the previous temporal information of the pollutants on yesterday to the model. To optimize the prediction model, the variable importance in the projection (VIP) of the partial least squares (PLS) is used to select key input variables as a preprocessing step. The prediction models are applied to a real indoor air quality dataset from telemonitoring systems data (TMS), which exhibits some nonlinear dynamic behaviors show that the selected key variables have strong influence on the prediction performances of the models. It demonstrates that the RNN model has the ability to model the nonlinear and dynamic system, and the predicted result of the RNN model gives better modeling performance and higher interpretability than other data-driven prediction models.

Key words: Air Quality Prediction, Nonlinear Modeling, Recurrent Neural Networks (RNN), Predicted Model, Partial Least Squares (PLS), Subway Station

INTRODUCTION

Metro systems, or underground or subway systems, have been considered as an important model of transport in order to enhance the quality of the transport, relieve congestion, as well as to fill gaps of insufficient public transport. The sort of harmful air pollutants are replaced and stayed in indoors such as subway stations which are used by many people. Also, serious indoor air pollution is caused by inadequate ventilation systems. Korea ministry established indoor air quality regulations to control major pollutants such as ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and PM (particulate matter) <10 μm in diameter (PM_{10}) or <2.5 μm in diameter ($PM_{2.5}$), due to their harmful effects on human health and vegetation. PM_{10} and $PM_{2.5}$ are potentially toxic to humans and all living matter [1].

To control the indoor air quality in subway systems, several key air pollutants are measured and monitored offline and also online by a telemonitoring (TMS) system which is built for the management of air quality in a subway station. To satisfy operational regulation for quality, safety and environmental constraints with minimum cost, the current status in a system should be measured and predicted. Due to increasing constraints and the necessity of a reliable environment, efficient modeling and monitoring methods are becoming more and more important. It is necessary to maintain the system performance as close as possible to optimal conditions, since the precise prediction of the air pollutant concentration and early fault detection in the subway environment are very critical to executing corrective action well before a dangerous situation occurs [1,2].

Fig. 1 shows an example of the air quality management in a subway station. First, multivariate monitoring based on principal component analysis (PCA) is used to diagnose the nine air pollutants in a TMS simultaneously from the viewpoint of ecological toxicology. Second, a contribution plot is used to identify and isolate the sources of bad or contaminated, or emergency air quality. Third, under the process information obtained from statistical monitoring techniques,

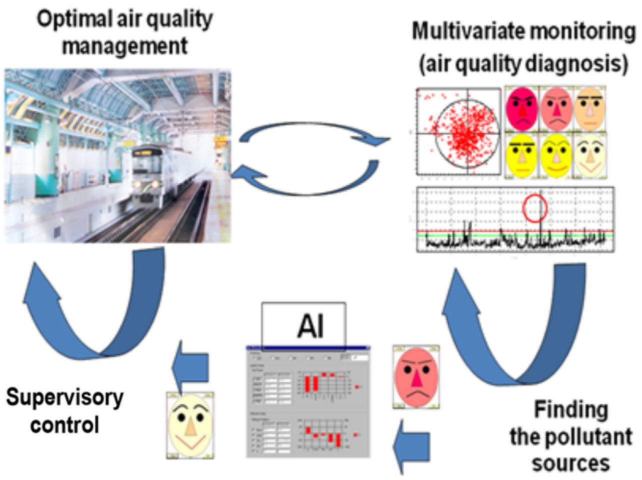

Optimal air quality management
Multivariate monitoring (air quality diagnosis)
AI
Supervisory control
Finding the pollutant sources

Fig. 1. An example of the air quality management in a subway station.

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

artificial intelligence (AI) such recurrent neural network, fuzzy, can be utilized to design the supervisory control algorithm and finally to optimize the operating condition. Here, the integrated monitoring of air quality can be accomplished by combining multivariate monitoring and the prediction method to check their toxicological health effects using statistical regression. In this paper, we focus on the prediction model of the air quality management.

Today, empirical data-based modeling is a widely used alternative to mechanistic modeling since it requires less specific knowledge of the process being studied compared to a first principles model. Here, data-driven empirical modeling techniques require data (measurements) which are collected on those variables believed to be representative of the process behavior and of the properties of the product or system output. The recurrent neural network (RNN), which is one of the data-driven models, recently has been applied successfully in various fields including environmental engineering, such as a modeling technique for nonlinear systems [3-5].

In this paper, a data-driven prediction method based on the RNN is developed to predict the air pollutants of the platforms in a subway station. Also, feed-forward neural networks (NN) and multiple regression models are used for comparisons. The proposed method has a preprocessing step of the partial least squares method. It can select the key variable to determine the structures of the RNN and other prediction models and interpret the variable relationship between input variables and select the key temporal information by its variable importance in the projection (VIP), and finally, it can predict the air pollutant of the platforms in a subway station through its soft sensing.

The outline of this paper is as follows. The first section introduces the basics of data-driven models of multiple linear regression, neural network, and recurrent neural network methods briefly and the proposed method is suggested. Results and discussion section contain illustrative application results in a real subway station. Finally, the conclusions of this article are addressed.

METHODS

1. Multiple Regression Model

Regression modeling is one of the techniques for predicting process design, optimization, process control, and other engineering activities. The regression model describes the relationship between independent and dependent variables. In the simple linear regression model, the dependent variable, or response, is related to one independent variable, or regressor. However, there are many empirical model building situations in which there are more than one independent variable [6].

A regression model that contains more than one independent variable is called multiple regression model. A multiple regression model includes interaction and quadratic effects. Depending on the values of the regression coefficients, the second-order model with interaction is capable of assuming a wide variety of shapes and thus is a very flexible regression model [7]. A multiple regression model which has more than two independent variables is represented by the following Eq. (1).

$$Y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_{12}x_1x_2+\beta_{11}x_1^2+\beta_{22}x_2^2+\varepsilon \tag{1}$$

where x_s are independent variables, Y is dependent variable, β_s are regression coefficients, and ε is random error term [6,7]. The goal of MLR is to find an approximation function for the prediction future response of the system outputs, which is to find a suitable approximation for the true functional relationship between independent variables and the response variables.

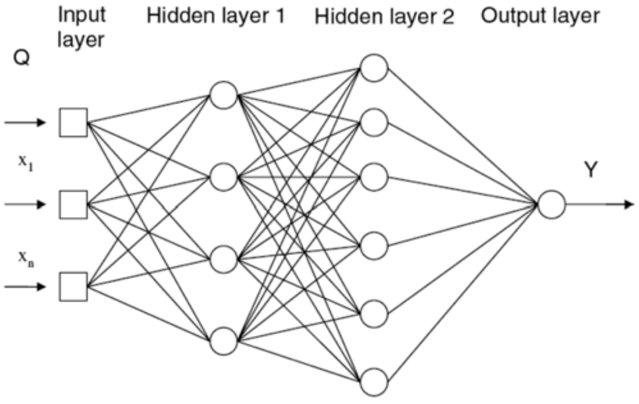


Fig. 2. The structure of multi-layer neural networks.

2. Neural Networks (NN) Model

The neural networks (NN) modeling in which the important operational features of the human nervous system are simulated can be applied to the modeling of complex nonlinear dynamic systems. The NN is a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes. Fig. 2 shows the structure of multi-layer NN. Multi-layer feed-forward NN used in this study consists of several layers: input, hidden layers 1 and 2, and an output layer, and each layer is comprised of several operating neurons [8]. Each neuron is connected to every neuron in adjacent layers before being introduced as input to the neuron in the next layer by connection weight, which determines the strength of the relationship between two connected neurons. Each neuron sums all the inputs that it receives, and the sum is converted to an output value based on a predefined activation, or transfer function. The sum of x_i (i=1, 2, …, n) multiplied with the corresponding weight factor w_i and critical value, b formed the neuron output y through transformed function, f following Eq. (2) [5,9].

$$y=f(net)$$

$$net=\sum_{i=1}^{n}x_iw_i+b \tag{2}$$

where n is the number of neurons of input, hidden and output layers, y is model output, f is transformed function, w_i is weight and b is bias.

To teach associations between inputs and outputs in a multiplayer perceptron with arbitrary hidden layers, the error backpropagation algorithm is used. Input vectors and the corresponding output vectors are used to train a network until input vectors are approximate in an objective function as defined by the user. The backpropagation algorithm refers to the manner in which the gradient is computed for nonlinear multiplayer networks. After the training stage as described above, properly trained backpropagation networks tend to give reasonable prediction values in the test stage. This generali-

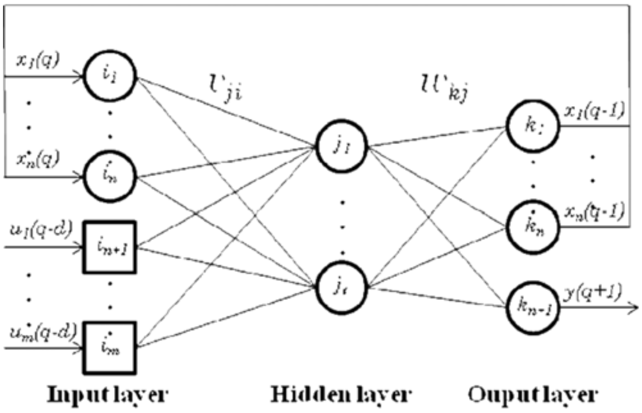


Fig. 3. The structure of a recurrent neural network.

zation property makes it possible to train a network on a representative set of input-output pairs and get good results without training the network on all the possible input-output data.

3. Recurrent Neural Networks (RNN) Model

The NN can be grouped into two categories: the feed-forward NN in which neurons have no loops, and the recurrent neural networks (RNN) in which loops occur because of feedback connections [5]. The feed-forward NN's' output depends on what is present at only the input layer. The RNN represents temporal meanings because of involving the previous states as well as current states [3,4]. Due to the dynamic nature of the recurrent neural network, it is capable of providing accurate one-step and moreover multi-step ahead predictions.

As shown in Fig. 3, an RNN consists of internal states, inputs, outputs, weights, activation functions, and feedback links. The current states are determined by the previous states, weights, and inputs as the following Eqs. (3), (4), and (5).

$$\hat{h}_j = f\left(\sum_{i=1}^{n} v_{ji}\hat{x}_i(q) + \sum_{i=n+1}^{n+m} v_{ji}u_{i-n}(q-d)\right) \tag{3}$$

$$\hat{x}_k(q+1) = \sum_{j=1}^{n} w_{kj}\hat{h}_j(q+1) \tag{4}$$

$$\hat{y}(q+1) = \sum_{j=1}^{i} w_{n+1j}\hat{h}_j(q+1) \tag{5}$$

where f is sigmoid function, v_{ji} and w_{kj} are weights between input layer and hidden layer, and hidden layer and output layer, respectively. The steepest gradient method is used for the learning.

4. Key Variables Selection of the Prediction Models

The data set in an environmental system is characterized by multivariate characteristics as well as collinear. To solve this problem, the first step of the prediction method is to extract the fundamental features of the data set, and the second step is to make a model for the data. In this paper, the PLS model is developed to select the key independent variables (**X**) to the response variables (**Y**), where response variables are the particulate matter quantity at the platform [10]. Here, the information of the PLS model is used for the variable selection of RNN model. PLS method can find the key process variables which are more influential on the response variables. Note that after the PLS weight vectors are computed, input variables are selected via the variable importance in the projection (VIP) of PLS, which is defined as following Eq. (6).

$$\mathrm{VIP} = \sum_{a}(\mathbf{w}_{ak})^2 \tag{6}$$

Suggested by [10], the VIP is calculated from the weight vector of the PLS model and the percentage that is explained by the dimension of the model. Thus, important inputs based on the VIP value can be selected, since the VIP is the sum over all model dimensions of the contributions. The VIP is a good measure of the influence of all input variables in the model on the response variables, such as the future concentrations of the particulate matters (PM_{10} and $PM_{2.5}$).

5. Data-driven Prediction Model of Indoor Air Quality by the Preprocessed Recurrent Neural Networks

Fig. 4 shows the proposed data-driven prediction models of the prediction of indoor air quality in a subway by the preprocessed recurrent neural networks. First, data of the air pollutants are collected in a subway station to develop a data-driven prediction model as a soft sensing. It has a preprocessing step, since too many variables have a bad effect on the prediction model. Second, the optimal input variables are determined by using the VIP of the PLS model. The VIP values are used for selecting the key input variables for the prediction models. Third, data-driven models for the indoor air quality are developed to predict the unknown future pollutant as a soft sensing model. Here, the RNN model predicts the air pollutant concentration at a platform in a subway station by using the previous information of PM_{10} at outside, temperature, humidity, the number of passengers and PM_{10} and $PM_{2.5}$ on yesterday. Two models of multiple regression model and neural network are used for the comparisons. Finally, the prediction results are compared with other prediction methods using the root mean square error (RMSE) criteria by following Eq. (7) between predicted and actual data.

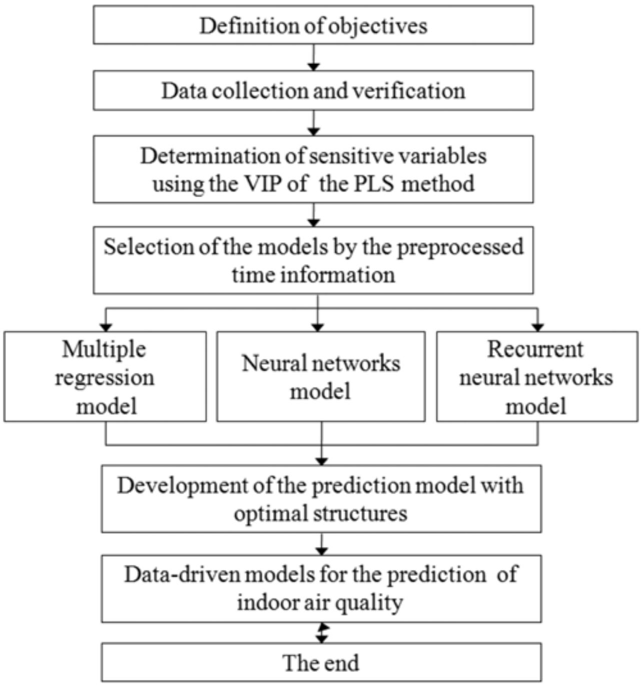

Definition of objectives
Data collection and verification
Determination of sensitive variables using the VIP of the PLS method
Selection of the models by the preprocessed time information
Multiple regression model
Neural networks model
Recurrent neural networks model
Development of the prediction model with optimal structures
Data-driven models for the prediction of indoor air quality
The end

Fig. 4. The proposed data-driven prediction models of the prediction of indoor air quality in a subway.

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(X_i-Y_i)^2}{n-1}} \tag{7}$$

RESULTS AND DISCUSSION

1. Monitoring Systems

There are two subway systems in Seoul, where the Seoul Metro Subway Corporation (SMSC) operates subway lines 1 to 4 and the Seoul Metropolitan Rapid Transit Corporation subway operates lines 5 to 8 [1]. We examined air pollutant data from a real-time TMS installed in four subway stations located on subway line numbers 1, 2, and 4 in Seoul. TMS systems are located at the center of the platform in each station and measure the concentration levels of seven air pollutants; NO, NO_2, NO_X, PM_{10}, $PM_{2.5}$, CO, CO_2, and temperature and humidity with the fixed measurement intervals. Fig. 5 shows the TMS at subway station B.

NO, NO_2 and NO_X concentrations were measured by the chemiluminescence of nitro-oxides materials and ozone, and PM_{10}, $PM_{2.5}$ concentrations were measured by the beta-ray attenuation principle with the corresponding size distribution filters. CO and CO_2 concentrations were measured by the non-dispersive infrared (IR) radiation absorption by CO and CO_2 molecules at the specific wavelengths. Meteorological data such as temperature and relative humidity strongly influence the efficiency of photochemical processes leading to noisy measurements and the formation of subsidiary particulates reflected in PM_{10} and $PM_{2.5}$ measurements [11].

In this study, the daily mean values for each variable from February 2007 to July 2008 with a total number of 490 observations are used. The first 358 observations were used to develop the prediction models. During this period, several observations indicative of abnormal air quality levels were omitted in order to ensure that the training data represented a normal condition, which was confirmed by the expert knowledge of an operator and a researcher. The remaining 132 observations were used as a test data set in order to verify the proposed method.

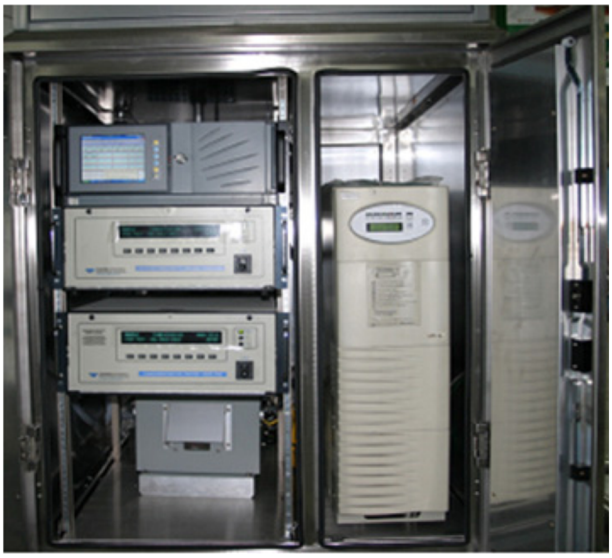


Fig. 5. Telemonitoring system (TMS) at B-subway station used in this study.

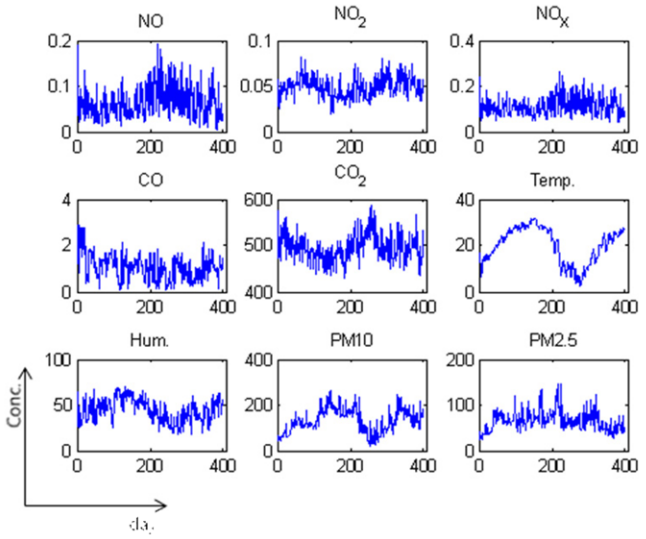


Fig. 6. Time series plot of nine air pollutants in a TMS system.

Fig. 6 shows the time series plot of nine air pollutants from the TMS at B-subway station. Here, the system monitors whether the concentration of a single pollutant is in or out of control, compared to the standard (regulation limit) determined by the Ministry of Environment. Although several peaks are observed, the concentrations of nitrogen components (NO_X, NO and NO_2) and particulate matter (PM_{10} and $PM_{2.5}$) are all technically within the standard limits of MOE.

To consider hydraulic characteristics, particulate matters (PM_{10} and $PM_{2.5}$) on yesterday are used for independent variables with the concentrations of nitrogen components (NO, NO_2 and NO_X), the concentrations of carbon components (CO and CO_2) two meteorological variables of temperature and humidity. And the concentrations of particulate matters (PM_{10} and $PM_{2.5}$) are used as the response variables. To extract the key input variables, the VIP plot which is determined by the importance of input variables is used in Fig. 7. Because the VIP can be considered as a measure of how much a certain influent variable corresponds to the samples, we are able to find important variables of the prediction model based on the VIP values. Here, PM_{10} and $PM_{2.5}$ on yesterday and temperature are selected as sensitive parameters, since they are more influential on the prediction than any other variables.

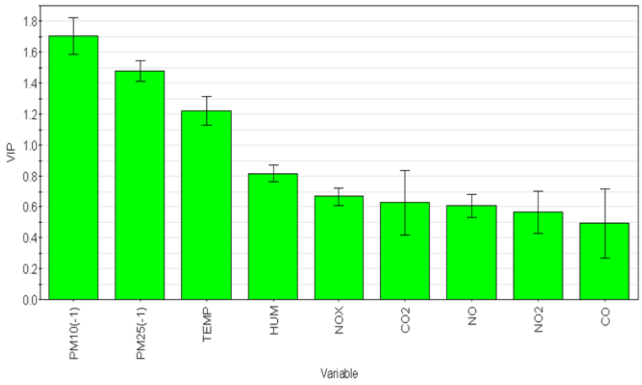


Fig. 7. PLS results as VIP plot.

The MLR, NN, and RNN models are developed to predict PM_{10} and $PM_{2.5}$ in a subway station as the following two case studies.

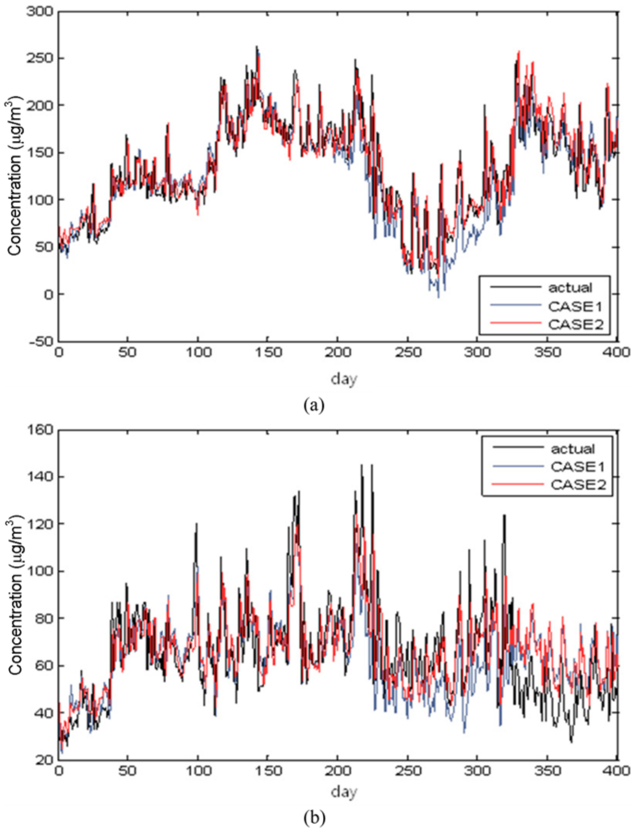
- Case 1: NO, NO_2, NO_X, CO, CO_2, Temperature, humidity, $PM_{10,t-1}$ and $PM_{2.5,t-1}$
- Case 2: $PM_{10,t-1}$, $PM_{2.5,t-1}$, and NO_X

All independent variables including the concentration of particular matters yesterday are used in case 1 and three sensitive parameters selected by the preprocessed PLS model are used in case 2.

Table 1 shows the RMSE values of three prediction models with MLR, NN, and RNN, which shows the regression performance of all models. RMSE values of the validation data set indicate that the error values not explained by the linear models are larger than nonlinear methods and the best prediction performance is achieved by the RNN model.

Table 1. The RMSE values of three prediction models with MLR, NN, and RNN

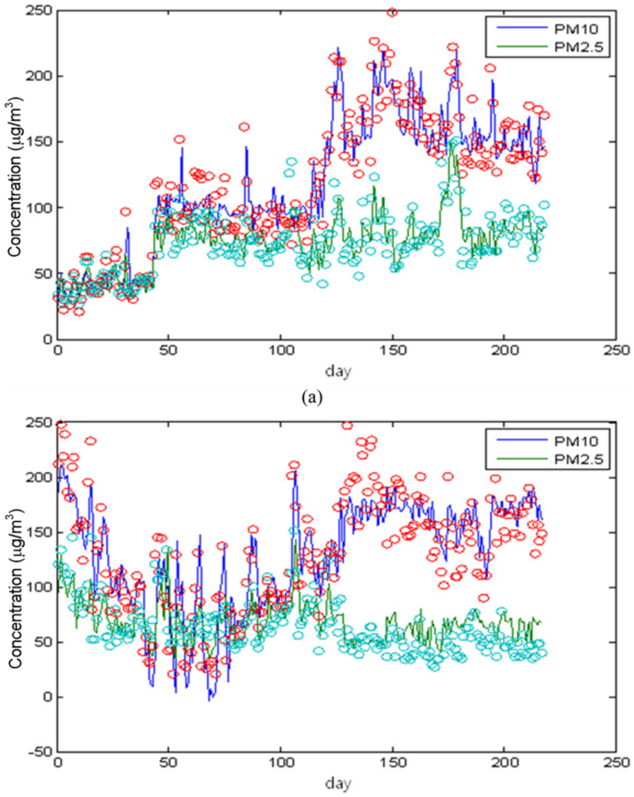
		PM_{10}		$PM_{2.5}$	
		Train.	Test	Train.	Test
Regression	Case 1	19.35	35.61	12.22	20.74
	Case 2	20.58	31.25	12.70	18.06
NN	Case 1	15.83	40.39	9.69	25.81
	Case 2	19.93	32.46	13.3	20.35
RNN	Case 1	18.65	29.37	12.01	18.38
	Case 2	20.64	28.57	13.25	17.80



(a)

(b)

Fig. 8. The prediction result of the multiple regression model for (a) PM_{10}, (b) $PM_{2.5}$.



(a)

(b)

Fig. 9. The prediction result of the NN model for (a) training data, (b) test data.

Figs. 8-10 show the predicted results of PM_{10} and $PM_{2.5}$ concentrations using three models. RMSE values of the MLR model for training data and test data in Fig. 8 are 20.58 and 31.25 on PM_{10} and 12.70 and 18.06 on $PM_{2.5}$, respectively, in case 2. The NN model in Fig. 9 consists of 9 or 3 variables in input layer, two hidden layers and 2 variables in output layer, where each hidden layers has 5 hidden nodes. Figs. 9(a) and (b) show the results for training and test data. RMSE values of PM_{10} and $PM_{2.5}$ of training data are 19.93 and 13.3, and test data are 32.46 and 20.35, respectively. The RNN model consists of 9 or 3 variables in input layer, a hidden layer, one variable in output layer and feedback connections, which was decided by the trial and error criteria. Figs. 10(a) and (b) show the results of training and test data for PM_{10} and Figs. 10(c) and (d) show the results for $PM_{2.5}$. RMSE values for training and test data are 20.64 and 28.57 on PM_{10}, 13.25 and 17.80 on $PM_{2.5}$, respectively, in case 2.

Overall, case 2 shows better prediction results of PM_{10} and $PM_{2.5}$ than case 1. It confirms that case 1 has unnecessary parameters and some input variables have a negative effect on the model prediction. In both PM_{10} and $PM_{2.5}$ prediction, the RMSE values of the RNN model are the lowest among the three prediction models. The RNN model shows a more accurate prediction capability than the other methods. The differences of RNN from NN and regression model are the time-delayed feedback links. This feature represents temporal meanings because of involving the previous states as well as current states. So, the predicted result from RNN shows the better

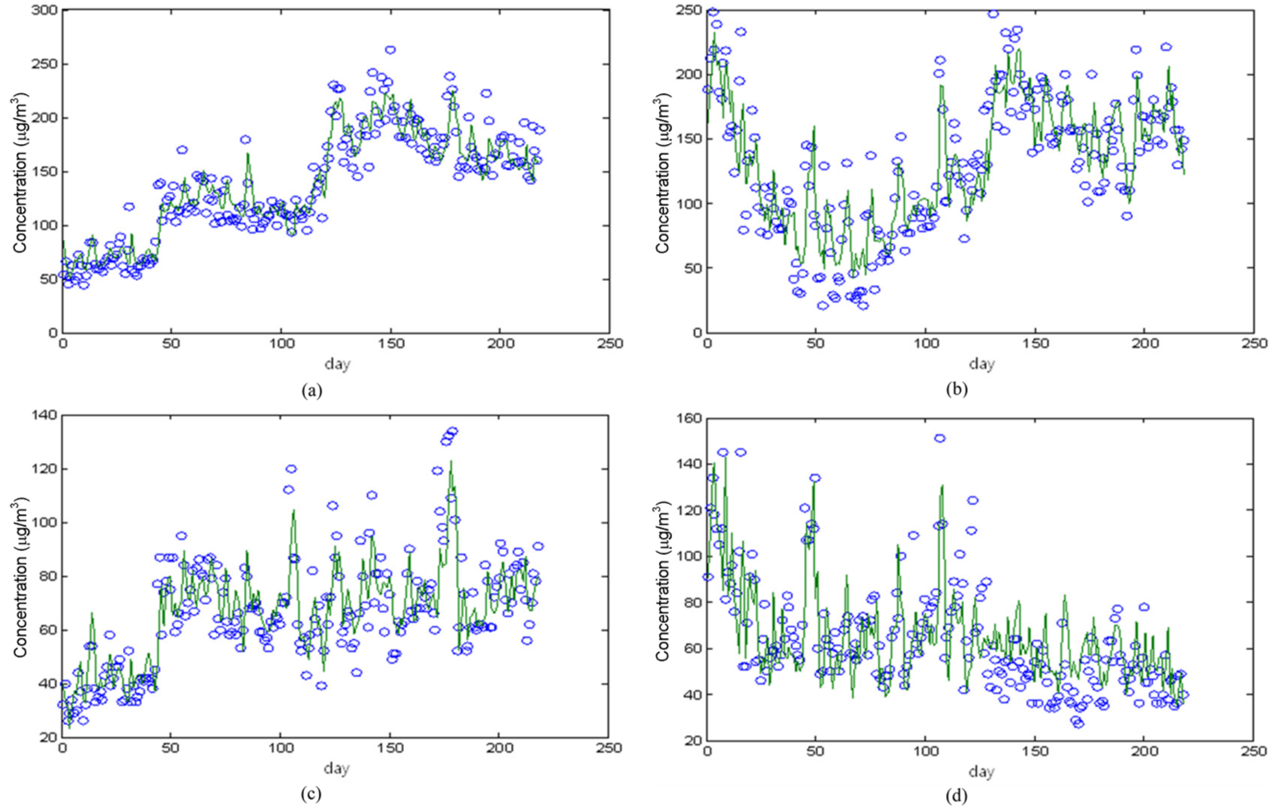


Fig. 10. The prediction result of the RNN model for (a) training data and (b) test data of PM_{10}, and (c) training data and (d) test data of $PM_{2.5}$.

modeling performance.

CONCLUSIONS

A data-driven prediction model with recurrent neural networks is proposed and compared with other prediction models of MLR and NN. To raise the efficiency of prediction, some additional input variables considering a system dynamics are added to the model input variables. Also, the VIP of the PLS is used to select the key input variables for the optimal structure of the models. The study shows that having too many variables has a bad effect on the prediction model. The RNN model by the preprocessed scheme shows better accurate prediction capability and has a lower RMSE values than other prediction models because it contains dynamic information of the previous pollutant concentration. This study confirms that RNN is more appropriate for the building of long-range prediction models of indoor air pollution process which has a large number of input and output terms in a system.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 20100001860) and the Seoul R&BD Program (CS070160).

REFERENCES

1. 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 KOSAE*, 571 (2006).
2. M. J. Nieuwenhuijsen, Levels of particulate air pollution, its elemental compositions, determinants and health effects in metro systems, *Atmospheric Environ.*, 7995 (2007).
3. S. H. Jung, *Improved recurrent neural networks for grammatical inference*, Ph. D thesis, KAIST, Korea (2000).
4. Y. D. Pan, S. W. Sung and J. H. Lee, *Control Eng. Practice*, **9**, 859 (2001).
5. T. Y. Pai, *Environ. Eng. Sci.*, **25**(5), 757 (2008).
6. D. C. Montgomery, G. C. Runger and N. F. Hubele, *Engineering statistics- 3rd Ed.*, WILEY, USA (2004).
7. R. E. Walpole, R. H. Myers, S. L. Myers and K. Ye, *Probability & statistics for engineers & scientists-8th Ed.*, Pearson Education International, USA (2007).
8. D. M. Himmelblau, *Ind. Eng. Chem.*, **47**, 5782 (2008).
9. D. J. Choi and H. K. Park, *J. Korean Soc. Water Quality*, **17**(1), 87 (2001).
10. D. V. Nguyen and D. M. Rocke, *Bioinformatics*, **18**(9), 1216 (2002).
11. D. M. Markovic, D. A. Markovic, A. Jovanovic, L. Lazic and Z. Mijic, *Environ. Monit. Assess.*, **145**, 349 (2008).