# AIR POLLUTION PREDICTION SYSTEM
# USING DEEP LEARNING

THANONGSAK XAYASOUK & HWAMIN LEE
Department Computer Software Engineering, Soonchunhyang University, South Korea

## ABSTRACT

One of the most influential factors on human health is air pollution, such as the concentration of $PM_{10}$ and $PM_{2.5}$ is a damage to a human. Despite the growing interest in air pollution in Korea, it is difficult to obtain accurate information due to the lack of air pollution measuring stations at the place where the user is located. Deep learning is a type of machine learning method has drawn a lot of academic and industrial interest. In this paper, we proposed a deep learning approach for the air pollution prediction in South Korea. We use Stacked Autoencoders model for learning and training data. The experiment results show the performance of the air pollution prediction system and model that proposed.
Keywords: *fine dust, $PM_{10}$, $PM_{2.5}$, air pollution prediction, deep learning.*

## 1 INTRODUCTION

One of the most influential factors on human health is air pollution, such as the concentration of $PM_{10}$ and $PM_{2.5}$ is a damage to a human. Hence, information about air quality is a great importance to protect human health and control air pollution. In May 2016, Korea's air quality was ranked 173 out of 180 countries, and the ultra-fine dust exposure was ranked 174 out of 180 countries. Particularly, fine dusts contain a large number of carcinogens, thereby increasing the incidence of various diseases. Therefore, the fine dust is regarded as the first risk factor of public health concern of the Korean people. According to data from the Max Planck Institute in Germany, the global death toll from air pollution is about 3.9 million per year, which is higher than the death rate from AIDS and malaria. According to the World Health Organization (WHO), 38,000 people die annually from air pollution in Korea.

However, at present, Korea has limited air pollution measurement stations such as fine dust and ozone, so it cannot obtain accurate air pollution information of the place where it is desired or the activity place. Fig. 1 shows the air pollution monitoring station of South Korea. Deep learning is a type of machine learning method has drawn a lot of academic and industrial interest [1]. In this paper, we propose a deep learning-based method for fine dust prediction. We use the deep learning algorithm to build a spatiotemporal prediction framework, which considers temporal and spatial relations of the dataset in modelling process. This model can predict the fine dust of all cities in South Korea and shows an accuracy of the process.

This paper is organized as follows. The related works are given in Section 2. Methodology is presented in Section 3. The experiments and results are conducted in Section 4. Finally, some conclusions will be shown in Section 5.

## 2 RELATED WORKS

A novel of spatiotemporal deep learning (STDL)-based air quality prediction method [2] considers spatial and temporal correlations are proposed. The stacked autoencoders (SAE) model is used to extract inherent air quality features, and it is trained in a greedy layer-wise manner. In [3], the stacked autoencoder model is used to learn generic traffic flow features and it also trained in a greedy layer-wise fashion.
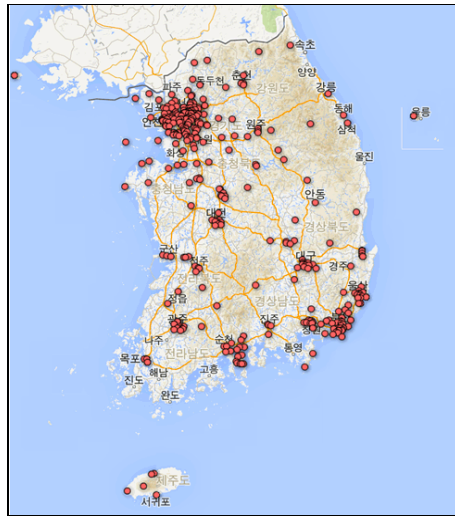
Figure 1:  The air pollution monitoring station of South Korea.

A multitask learning (MTL) [4], a multi-task regression layer is used above the Deep Belief Networks for supervised prediction and investigate homogeneous MTL and Heterogeneous MTL for prediction.

## 3  METHODOLOGY

The stacked autoencoder model is a stack of autoencoders, which is a famous deep learning model. It builds a block to construct a deep network [5].

### 3.1  Autoencoder

The autoencoder is a neural network that attempts to reproduce its input, i.e., the target output is the input of the model. Fig. 2 gives an illustration of autoencoder, which has one input layer, one hidden layer, and one output layer.
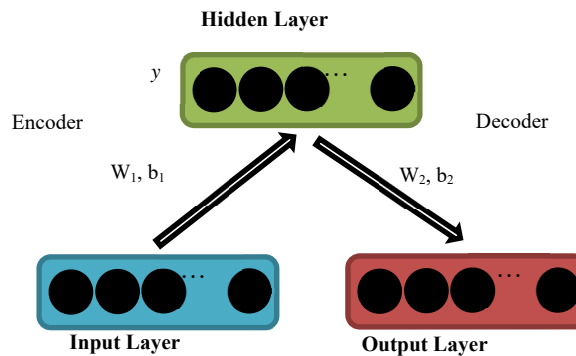


Figure 2:  Autoencoder model.

Given a set of training to find the equation to get samples, where $x^{(i)} \in R^d$, an autoencoder first encodes an input $x^{(i)}$ to a hidden representation $y(x^{(i)})$ based on (1), and then it decodes representation $y(x^{(i)})$ back into a reconstruction $z(x^{(i)})$ computed as in (2), as shown in

$$y(\mathrm{x}) = f(\mathrm{W}_1 x + b), \tag{1}$$

$$z(\mathrm{x}) = g(\mathrm{W}_2 y(\mathrm{x}) + c), \tag{2}$$

where $W_1$ is a weight matrix, $b$ is an encoding bias vector, $W_2$ is a decoding matrix, and c is a decoding bias vector; we consider logistic sigmoid function $1/(1 + \exp(-x))$ for f(x) and g(x) in this paper. By minimizing reconstruction error L(X, Z), we can obtain the model parameters, which are here denoted as θ, as

$$\theta = arg_\theta \min L(X,Z) = arg_\theta min \frac{1}{2}\sum_{i=1}^{N}\left\|x^{(i)} - z(x^{(i)})\right\|^2. \tag{3}$$

One serious issue concerned with an autoencoder is that if the size of the hidden layer is the same as or larger than the input layer, this approach could potentially learn the identity function. However, current practice shows that if nonlinear autoencoders have more hidden units than the input, restrictions such as sparsity constraints are imposed, this is not a problem. When sparsity constraints are added to the objective function, an autoencoder becomes a sparse autoencoder, which considers the sparse representation of the hidden layer. To achieve the sparse representation, we will minimize the reconstruction error with a sparsity constraint as

$$SAO = L(X,Z) + \gamma \, dfsd \sum_{j=i}^{N} KL(\rho||\hat{\rho}_j), \tag{4}$$

where $\gamma$ is the weight of the sparsity term, $H_D$ is the number of hidden units, $\rho$ is a sparsity parameter and is typically a small value close to zero, $\hat{\rho}_j = (1/N)\sum_{i=1}^{N} y_j(x^{(i)})$ is the average activation of hidden unit j over the training set, and $KL(\rho \parallel \hat{\rho}_j)$ is the Kullback–Leibler (KL) divergence, which is defined as

$$KL(\rho \parallel \hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1 - \rho)\log \frac{1-\rho}{1-\hat{\rho}_j}. \tag{5}$$

The KL divergence has the property that $KL(\rho\hat{\rho}_j) = 0$ if $\rho = \hat{\rho}_j$. It provides the sparsity constraint on the coding. The backpropagation (BP) algorithm can be used to solve this optimization problem.

### 3.2  Stacked autoencoders

An SAE model is created by stacking autoencoders to form a deep network by taking the output of the autoencoder found in the layer below as the input of the current layer [5]. More clearly, considering SAEs with l layers, the first layer is trained as an autoencoder, with the training set as inputs. After obtaining the first hidden layer, the output of the kth hidden layer is used as the input of the (k + 1)th hidden layer. In this way, multiple autoencoders can be stacked hierarchically. This is illustrated in Fig. 3.
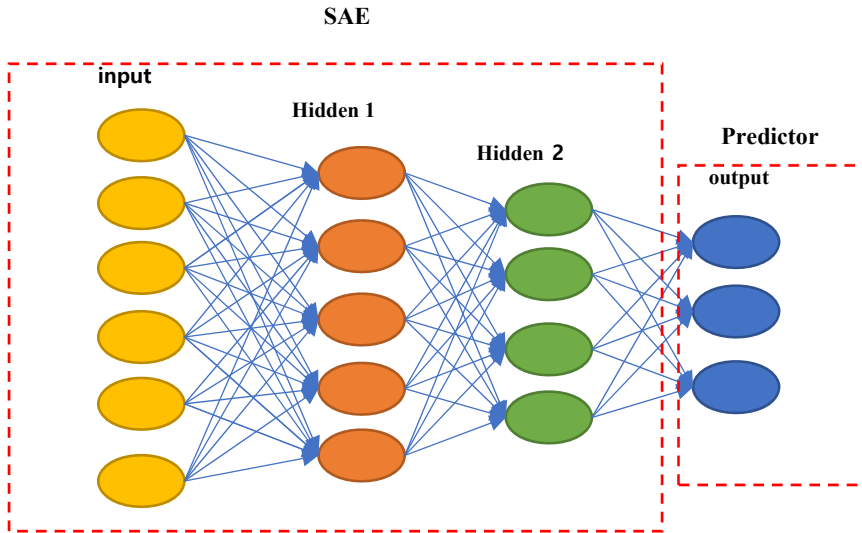
## 3.3  Training algorithm

The learning algorithm we used for fine dust prediction is shown in Algorithm 1.

## 4  EXPERIMENTS AND RESULTS

### 4.1  Dataset

In this paper, we use some dataset such as Table 1, the concentration of fine dust ($PM_{10}$ and $PM_{2.5}$), weather information (wind speed and direction, temperature, sky condition, rain, humidity, rain condition) and location information for 2017 in South Korea.



Figure 3:  Stacked autoencoder.

---

**Algorithm 1. Training SAEs**

---

Given training samples $X$ and the desired number of hidden layers $l$,

Step 1) Pre-training the SAE

  i.   Set the weight of sparsity $\gamma$, sparsity parameter $\rho$, initialize weight matrices and bias vectors randomly.
  ii.  Greedy layer-wise training hidden layers.
  iii. Use the output of the k*th* hidden layer as the $\{W_1^{k+1}, b_1^{k=1}\}_{k=0}^{l-1}$ and input of the (k + 1)*th* hidden layer. For the first hidden layer, the input is the training set.
  iv.  Find encoding parameters for the (k + 1)*th* hidden layer by minimizing the objective function.

Step 2) Fine-tuning the whole network

  i.   Initialize$\{W_1^{l+1}, b_1^{l+1}\}$randomly or by supervised training.
  ii.  Use the BP method with the gradient-based optimization technique to change the whole network's parameters in a top-down fashion.

---

Table 1: Our data model.

| Attributes | | Unit | Attributes | | Unit |
|---|---|---|---|---|---|
| Area | | City | Date | | yyyymmdd |
| Site code | | Number | Time | | hour |
| Temperature | | Celsius | Rain | | mm |
| Wind speed | | m/s | Humidity | | g/m$^3$ |
| Wind direction | (0–90) North | Degree | Rain condition | (0) No rain | |
| | (90–180) East | | | (1) Rainy | |
| | (180–270) South | | | (2) Sleet | |
| | (270–360) West | | | (3) Snow | |
| PM$_{10}$ | (0–30) Good | µg/$m^3$ | PM$_{2.5}$ | (0–15) Good | µg/$m^3$ |
| | (31–50) Normal | | | (16–25) Normal | |
| | (51–100) Bad | | | (26–50) Bad | |
| | (100 ~) Dangerous | | | (50 ~) Dangerous | |

We constructed the database using the fine dust and weather information data provided by the Air Korea website [4] and the Korea Meteorological Agency website [5]. We constructed a database of 695,041 data from January 1, 2017, to December 31, 2017. We defined data modeling as shown in Table 1 to analyze the characteristics of fine dust the correlation between fine dust and weather conditions.

In Table 1, we were selected twelve variables to be the input nodes. The input variables are area, data, time, site code, temperature, wind speed, wind direction, rain, rain condition, humidity, PM$_{10}$ concentration and PM$_{2.5}$ concentration. The output variable is the predicted value of fine dust (PM$_{10}$ or PM$_{2.5}$).

## 4.2 Prediction accuracy

To evaluate the performance of the proposed model, we adopted the performance index is the Root-Mean-Square error (RMSE). RMSE was calculated between the measured Air pollution value and predicted value by using the model. The index is defined as:

$$RMSE = \sqrt{\sum_{i=1}^{N}(P_m - P_r)^2/N}, \tag{6}$$

where $P_m$ denotes the measured air pollution, $P_r$ denotes the predicted air pollution a$nd$ $N$ $is$ the number of measured data.

Table 2:  Prediction results for each area.

| Area | RMSE PM$_{10}$ | RMSE PM$_{2.5}$ | Area | RMSE PM$_{10}$ | RMSE PM$_{2.5}$ |
|---|---|---|---|---|---|
| Busan | 14.64 | 7.79 | Incheon | 21.01 | 8.63 |
| Daegu | 9.08 | 7.39 | Sejong | 7.77 | 2.72 |
| Daejeon | 22.93 | 4.78 | Seoul | 5.71 | 7.97 |
| Gwangju | 5.16 | 2.18 | Ulsan | 8.97 | 10.17 |

### 4.3  Experiment results

Table 2 indicates that our model showed the different prediction results for each area in South Korea. Gwangju city has the best performance with the lowest RMSE (Root Mean Square Error) PM$_{10}$ value of 5.16 and the lowest RMSE PM$_{2.5}$ value of 2.18.

Fig. 4 and Fig. 5 show the comparison of test and predict graph of PM$_{10}$ and PM$_{2.5}$. The value of fine dust is plotted along the Y axis, and the month is plotted along the X axis. In the experimental results, we confirmed that the predicted values of fine dust through learning and actual data show a similar pattern. It is necessary to modify the model considering the time when the predicted pattern appears slightly later than the actual data.

From the result obtained of experiment, it can observe and compare with the Air Quality Index (AQI) in Table 1. In Fig. 4, a highest PM$_{10}$ concentration in Busan is 41.11 µg/$m^3$ in February, a highest PM$_{10}$ concentration in Daegu is 50.29 µg/$m^3$ in February, a highest PM$_{10}$ concentration in Daejeon is 71.17 µg/$m^3$ in October, a highest PM$_{10}$ concentration in Gwangju is 21.75 µg/$m^3$ in January, a highest PM$_{10}$ concentration in Incheon is 49.76 µg/$m^3$ in April, a highest PM$_{10}$ concentration in Sejong is 78.32 µg/$m^3$ in June, a highest PM$_{10}$ concentration in Seoul is 72.69 µg/$m^3$ in March, and a highest PM$_{10}$ concentration in Ulsan is 52.36 µg/$m^3$ in June. In Fig. 5, a highest PM$_{2.5}$ Concentration in Busan is 24.69 µg/$m^3$ in January, a highest PM$_{2.5}$ Concentration in Daegu is 38.43 µg/$m^3$ in February, a highest PM$_{2.5}$ Concentration in Daejeon is 51.56 µg/$m^3$ in October, a highest PM$_{2.5}$ Concentration in Gwangju is 16.36 µg/$m^3$ in January, a highest PM$_{2.5}$ Concentration in Incheon is 22.53 µg/$m^3$ in April, a highest PM$_{2.5}$ Concentration in Sejong is 42.26 µg/$m^3$ in May, a highest PM$_{2.5}$ Concentration in Seoul is 52.25 µg/$m^3$ in June, and a highest PM$_{2.5}$ Concentration in Ulsan is 36.53 µg/$m^3$ in February.

## 5  CONCLUSIONS

Deep learning approaches are increasing in their popularity every day. Deep learning provides fast and effective solutions, especially in the analysis of big data. In this paper, we introduced a deep- learning-based method for fine dust prediction and used the deep learning algorithm to build a spatiotemporal prediction framework, which considers temporal and spatial relations of the dataset in modelling process. We showed the experiment results of prediction and accuracy of our deep learning model.
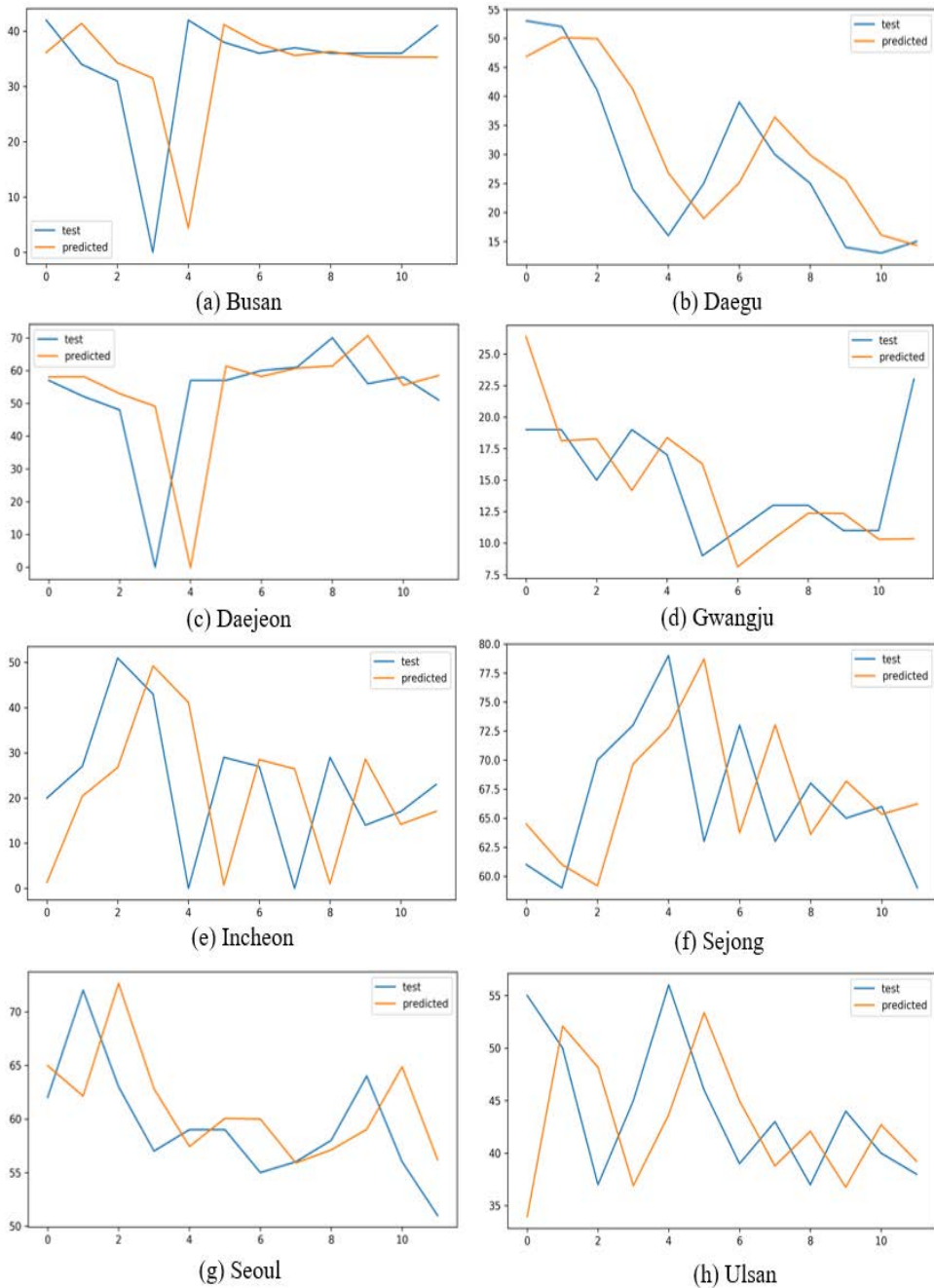
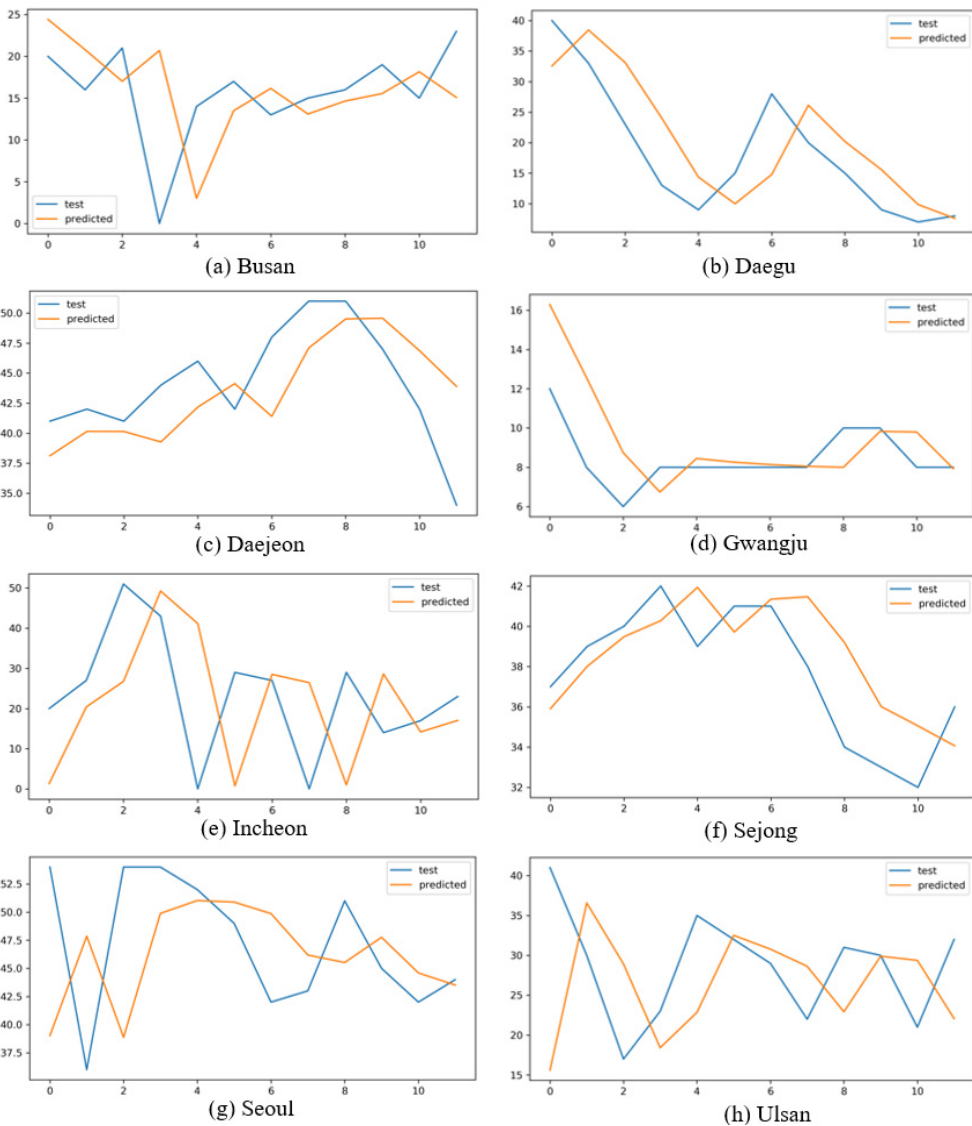Figure 4:  Comparison of actual data and predicted data – $PM_{10}$.

Figure 5: Comparison of actual data and predicted data – PM$_{2.5}$.

REFERENCES

[1] Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S. & Lew, M.S., Deep learning in neural network: An overview. *Neural Network*, **61**, pp. 85–117, 2015.

[2] Li, X., Peng, L., Hu, Y., Shao J. & Chi, T., Deep learning architecture for air quality predictions. *Environment Science and Pollution Research*, pp. 22408–22417, 2016.

[3] Lv, Y., Duan, Y., Kang, W., Li, Z. & Wang, F.-Y., Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, **16**(2), 2015.

[4] Huang, W., Song, G. Haikun H. & Kunqing, X., Deep architecture for traffic flow prediction: Deep Belief Networks with multitask learning. *IEEE Transactions on Intelligent Transportation Systems,* **15**(5), 2014.

[5] Bengio, Y., Lamblin, P., Popovici, D. & Laroachelle, H., Greedy layer wise training of deep networks. *Proceedings of Adv. NIPS*, pp. 153–160, 2007.

[6] Air Korea. www.airkorea.or.kr. Accessed on: 20 Oct. 2017.

[7] Weather data release portal. http://data.kma.go.kr. Accessed on: 20 Oct. 2017.