International Journal of Advanced Smart Convergence Vol.11 No.2 30-37 (2022)
http://dx.doi.org/10.7236/IJASC.2022.11.2.30

IJASC 22-2-5

Prediction of changes in fine dust concentration using LSTM model

Gi-Seok Lee*, Sang-Hyun Lee**

** Representative, Uclab Inc, Gwangju, Korea*
*** Associate Professor, Department of Computer Engineering, Honam University, Korea*
leegiseok@gmail.com, leesang64@honam.ac.kr

Abstract

Because fine dust (PM10) has a close effect on the environment, fine dust generated in the climate and living environment has a bad effect on the human body. In this study, the LSTM model was applied to predict and analyze the effect of fine dust on Gwangju Metropolitan City in Korea. This paper uses prediction values of input variables selected through correlation analysis to confirm fine dust prediction performance.

In this paper, data from the Gwangju Metropolitan City area were collected to measure fine dust. The collection period is one year's worth of data was used from january to December of 2021, and the test data was conducted using three-month data from January to March of 2022. As a result of this study, in the as a result of predicting fine dust (PH10) and ultrafine dust (PH2.5) using the LSTM model, the RMSE was 4.61 and the test result value was as low as 4.37. This reason is judged to be the result of the contents of the one-year sample.

Keywords: *LSTM model, Fine dust (PH10), Ultrafine dust (PH2.5), Correlation analysis, Fine dust concentration.*

1. Introduction

Particulate matter (PM) is a substance in the form of particles with various sizes, shapes, and components, and when the size per particle size is less than 10 μg, it means PM10, 2.5 μg((10-6g), and when it is less than PM2.5. Fine dust is an inhaled particle that is small enough to pass through the chest area of the respiratory tract, and if it contains heavy metals, it can seriously affect the human body. It can pose a greater risk to human health than other air pollutants such as ozone and carbon monoxide. As a result, many studies have been conducted on the effects of fine dust on the human body, and research results have been published that it can affect cardiovascular, respiratory, and cerebrovascular diseases [1-2]. In addition, according to the 2016 Organization for Economic Co-operation and Development (OECD) report, the premature mortality rate from outdoor fine dust and ozone in Korea was 1,109 per 1 million people, the highest among OECD countries. announced that it is visible [3].

According to the contents announced by Korea Environment Corporation in 2020, fine dust is primary aerosol, mainly composed of liquid/solid metal and carbon mixtures depending on the production process, and gaseous organic and inorganic compounds: nitrogen compounds (NOx), sulfur compounds (SOx), volatile

Manuscript Received: March. 14, 2022 / Revised: March. 17, 2022 / Accepted: March. 20, 2022
Corresponding Author: leesang64@honam.ac.kr
Tel: +82-62-940-5285, Fax: +82-62-940-5285
Associate Professor, Department of Computer Engineering, Honam University, Korea

organic compounds (VOCs), etc. formed through a photochemical reaction classified as a secondary pollutant (Secondary aerosol). Primary fine dust is emitted directly from fuel combustion facilities, and air pollution emissions can be calculated, but it is difficult to calculate the direct emission of secondary pollutants [4].

According to the Comprehensive Measures for Fine Dust Management Announced in September 2017 in Korea, secondary-generated substances account for about 72% of the total emissions on a nationwide basis, it was found that the amount of dust generated was the highest, and NOX and VOCS showed a high need for management as substances contributing to the production of O3.

In this paper, we analyze the correlation of major factors through correlation analysis and, through this, improve the precision of fine dust prediction. In addition, we intend to propose analysis and prediction by applying the LMST model of deep learning.

For the meteorological and air pollution factors used in the fine dust prediction model, we analyzed the correlation between fine dust and each factor by collecting historical data from Jan. 2021 to Dec. 2021 in the Gwangju Metropolitan City area, and analyzed aspects of PM10 and PM2.5. In this way, the characteristics of the data used in the predictive model are identified. In addition, input variables for each model to be used in the proposed predictive model are selected through the correlation analysis results.

2. Correlation Analysis of Meteorological and Air Pollutants and Fine Dust Concentration

The fine dust prediction model using deep learning learns the model using the past weather and air pollutants whose correlation with fine dust has been revealed as input variables. When an input variable with low correlation is used as the training data of the predictive model, there is a problem that the model cannot accurately find the pattern inherent in the data, and thus the prediction performance may deteriorate [5-6].

In order to "analyze the correlation between weather and air pollutants and fine dust concentrations", "weather and air pollutant data for the Gwangju Metropolitan City in 2021 were collected". "For meteorological data, the Open Meteorological Data Portal [8]" was used, and data on temperature, average wind speed, maximum wind direction, and relative humidity were collected. In the case of air pollutant data, using Air Korea [9], ozone (O3), nitrogen dioxide (NO2), carbon monoxide (CO), sulfur dioxide (SO2), fine dust concentration (PH10), ultrafine dust (PH2.5) were collected. For each data, final confirmed data by time of 2021 was used, and daily average data was used for air pollutant and fine dust concentration data and temperature data to minimize the effect of missing values.

The training data consists of 455 data in 2021 and 11 column values in Fig. 1, and the test data consists of 90 data from January to March 2022.

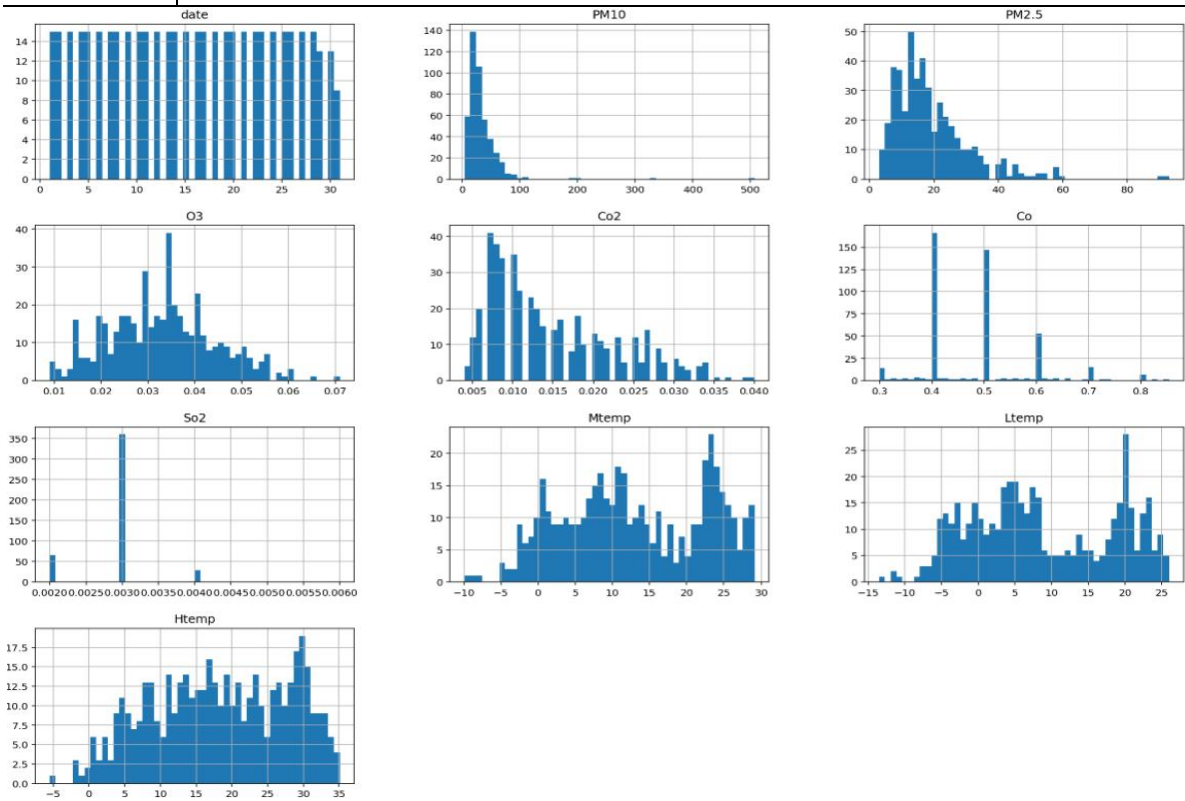
	year-mouth	date	PM10	PM2.5	O3	Co2	Co	So2	Mtemp	Ltemp	Htemp
0	2021-1	1	23	18	0.019	0.018	0.439	0.003	-0.6	-5.0	3.5
1	2021-1	2	33	21	0.023	0.014	0.449	0.003	-0.3	-3.2	2.6
2	2021-1	3	38	20	0.027	0.012	0.456	0.003	-0.6	-3.5	3.7
3	2021-1	4	48	33	0.010	0.033	0.665	0.003	1.2	-3.6	6.9
4	2021-1	5	37	28	0.020	0.021	0.534	0.003	-0.2	-4.0	3.5

Figure 1. Sample data for analysis of meteorological and air pollutants and fine dust concentrations

Table 1 and the graph show the correlation values to confirm the correlation of the variables used in this study.

Table 1. Correlation analysis table and graph to check the correlation of variables

Division	date	PM10	PM2.5	O3	Co2	Co	So2	Mtemp
Mtemp								1.00
So2							1.00	0.37
Co						1.00	0.19	0.18
Co2					1.00	0.74	0.06	0.52
O3				1.00	0.59	0.33	0.10	0.37
PM2.5			1.00	0.04	0.59	0.63	0.02	0.27
PM10		1.00	0.76	0.06	0.26	0.30	0.01	0.11
date	1.00	0.05	0.01	0.07	0.00	0.01	0.02	0.04



Generally, 0.7 or more is defined as a strong correlation, and 0.3 or more is defined as having a weak correlation, so that the relationship with variables greater than 0.3 is mainly identified. PM10 and PM2.5 were confirmed to have a strong correlation at 0.76, and PM2.5 and Co2 and O3 and Co2 were found to have a strong correlation with each other at 0.59. PM2.5 and Co had a correlation of 0.63, and Co2 and Co had a strong correlation of 0.75. Finally, Mtemp, So2 and O3 showed a weak correlation of 0.37, and Mtemp and Co2 showed a correlation of 0.52. Therefore, it was found that So2, Co, PM10, and PM2.5 were all related as they are air pollutants.

3. Composition of data set

Air pollutants and meteorological data in the Gwangju metropolitan area were collected based on the results of correlation analysis for the design and experiment of the fine dust prediction model. The total data collected is shown in Table 2. In the case of data, in order to obtain a recently generated data sample, 455 pieces of data for a year were collected from January 2021 to December 2021, of which 450 were used except for 5 missing

values, and the test data was From January to March 2022, 90 pieces of data were collected for a 3-month period.

Table 2. Number of training data

Category	Item	Site to receive the source	Number of data	Region
Air pollutants	PM10	https://www.airkorea.or.kr/index 2021.01.01~2021.12.31 (12 months)	455	Gwangju metropolitan area
	PM2.5		455	
	O3		455	
	Co2		455	
	Co		455	
	So2		455	
Temperature data	Mtemp	https://web.kma.go.kr/weather/forecast/timeseries.jsp	455	

Table 3. Number of test data

Category	Item	Site to receive the source	Number of data	Region
Air pollutants	PM10	https://www.airkorea.or.kr/index 2022.01.01~2022.03.31 (3 months)	90	Gwangju metropolitan area
	PM2.5		90	
	O3		90	
	Co2		90	
	Co		90	
	So2		90	
Temperature data	Mtemp	https://web.kma.go.kr/weather/forecast/timeseries.jsp	90	

4. Model design and Implementation

The was implemented using LSTM (Long short-term memory), a model of deep learning. LSTM is an algorithm that overcomes the long-term dependency problem of RNN (Recurrent Neural Network), and is equipped with a function to remember information for a long time. Basically, in RNN, the module with a loop has only one layer, whereas LSTM has the same structure and has 4 layers.

Table 4. Design of LSTM model

```
# Continue mapping the output to the input of the LSTM
# Selection of multi-layer perceptron model function for time series prediction
# Split the created array into multiple input/output patterns
# If 5 are input as input values, 1 is outputted.steps=5
features=1 # Acts as an input layer
def split_sequence(sequence, steps): # split a univariate sequence into samples
    x, y = list(), list()
# find the end of this pattern
    for i in range(len(sequence)):
        end_index = i+steps
# check if we are beyond the sequence
        if end_index > len(sequence)-1:
            break
# gather input and output parts of the pattern
```

```
        seq_x, seq_y=sequence[i:end_index], sequence[end_index]
        x.append(seq_x)
        y.append(seq_y)
    return array(x), array(y)
```

The most commonly used optimization algorithm in deep learning uses the Adaptive Moment Estimation (Adam) function. In order to use the Adam Optimization Algorithm, initialization must be performed first, and v and S used in Momentum and RMSprop must be performed as in Equation (1).

$$\begin{aligned} v_{dW} &= 0, S_{dW} = 0, \; v_{db} = 0, S_{db} = 0 \\ v_{dW} &= \beta_1 v_{dW} + (1-\beta_1) dW \\ v_{db} &= \beta_1 v_{db} + (1-\beta_1) db \\ S_{dW} &= \beta_2 S_{dW} + (1-\beta_2) dW^2 \\ S_{db} &= \beta_2 S_{db} + (1-\beta 2) db 2 \end{aligned} \tag{1}$$

Here, the bias correction introduced by Momentum is performed as in Equation (2).

$$\begin{aligned} v_{dw}^{biascorr} &= v_{dw} / (1- \beta^t) \\ v_{db}^{biascorr} &= v_{db} / (1- \beta^t) \\ s_{dw}^{biascorr} &= v_{dw} / (1- \beta^t) \\ s_{db}^{biascorr} &= s_{db} / (1- \beta^t) \end{aligned} \tag{2}$$

Finally, the weight update is performed using both the weight update method of Momentum and RMSprop.

$$\begin{aligned} \mathrm{W} &= \mathrm{W} - \alpha v_{dW}^{biascorr} / \sqrt{s_{dW}^{biascorr} + \epsilon} \\ b &= b - \alpha v_{dW}^{biascorr} / \sqrt{s_{dW}^{biascorr} + \epsilon} \end{aligned} \tag{3}$$

Here, Adam hyperparameters are as follows.

α : learning rate, $\beta 1$: first-order moment, mostly 0.9 (calculated exponentially weighted average of dw), $\beta 2$: Quadratic moment, 0.99 (calculated as exponentially weighted average of *dw2* and *db2*)

ϵ : In the paper 0.10^{-8} (Little to no performance impact)

For the implementation of this paper, *ReLU* was selected as the activation function, and learning was carried out using least squares error as the loss function.

The loss value was used to find the optimal parameter value, and in <Table 5> 500 epochs were performed to confirm the optimal value. As a result, the lowest value of less 0.9588 was output at 436 epoch, and the After that, the less value increased slightly again, so in this study, 436 epoch, which was output as the optimal value, was set and performed.

Table 5. Find the optimal loss value

Epoch	Loss
1	1957.2372
50	614.3231
100	157.7417
150	78.7454
200	17.0119
250	13.7876
300	5.2558
350	6.9767

400	2.3229
436	0.9588

In this paper, the model selected to improve the prediction performance was performed with LSTM, and 'Adam' was applied to optimize the *ReLU* function as the activation function used here.

Table 6. Selected model and applied function

```
# The selected model proceeded with LSTM, and 'Adam' was applied to optimize the reLU function and the activation function used here.
model=Sequential()
model.add(LSTM(300, activation='relu', input_shape=(steps, features)))
model.add(Dense(1))
model.compile(optimizer='adam', loss='mse')
histogram=model.fit(x_PM10, y_PM10, epochs=436, batch_size=100, verbose=1)
loss_and_metrics = model.evaluate(x_test, y_test, batch_size=32)
```

In order to improve the prediction performance of the LSTM models used in this paper, the model was designed through optimal hyper-parameter setting during model creation. The model was implemented in the learning environment of python, and the library used to implement the model was *keras* and *scikit-learn*.

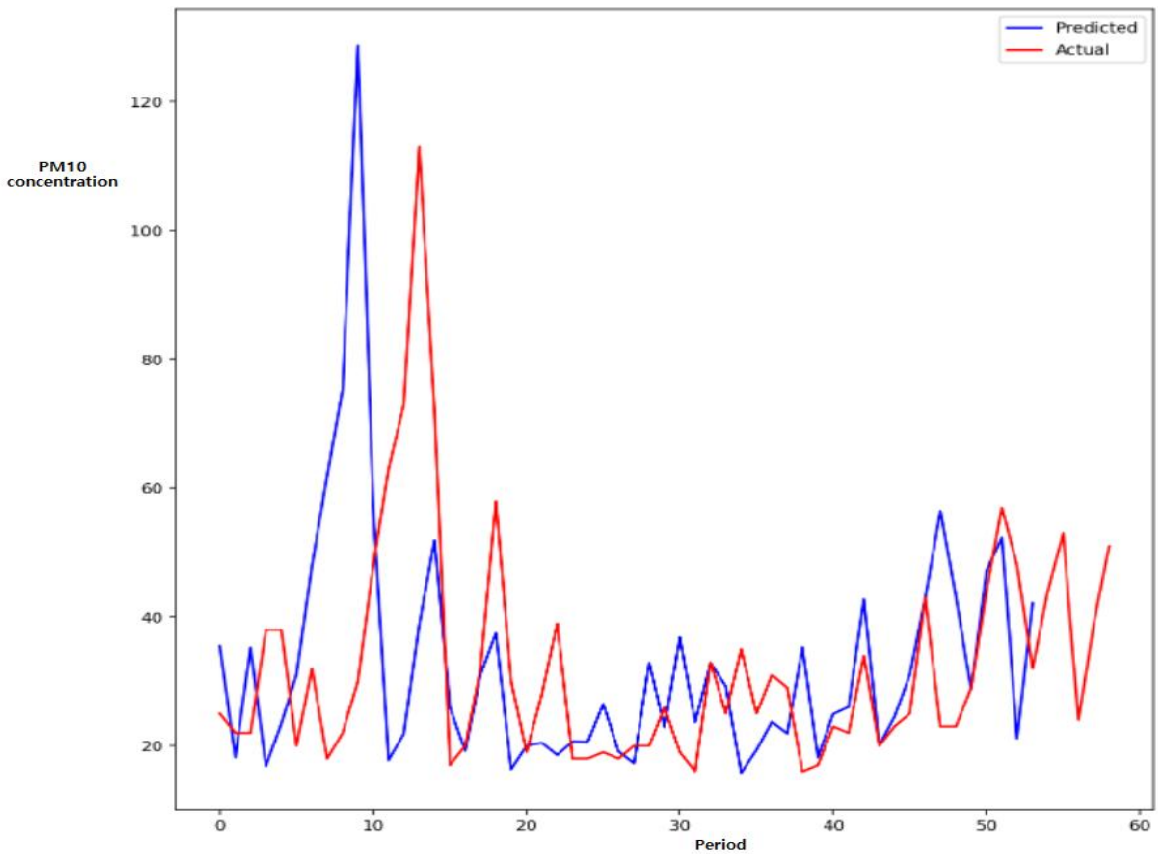


Figure 2. Execution result of fine dust (PH10) prediction model using LSTM model

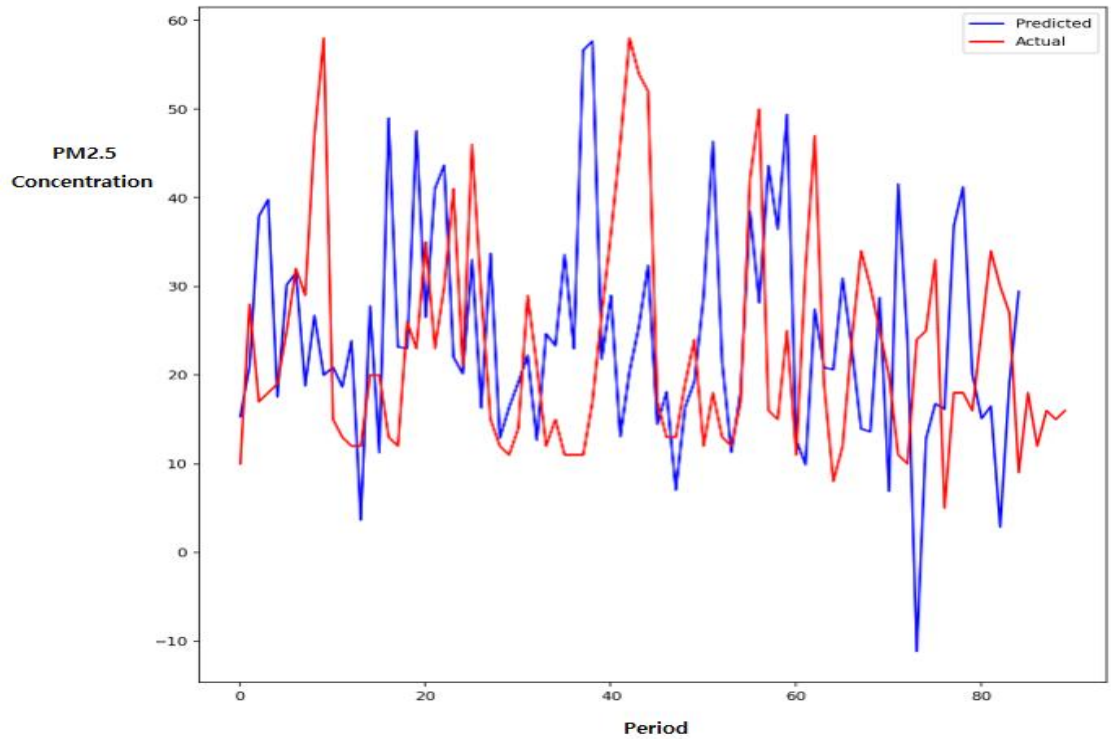


Figure 3. Execution result of fine dust (PH2.5) prediction model using LSTM model

Figure 2 shows the prediction results of fine dust (PH10) using the LSTM model, and Figure 3 shows the prediction execution results of ultrafine dust (PH2.5). As a result of learning, the training result was RMSE 4.61, and the test result value was 4.37.

5. Conclusion

In this paper, correlation analysis was performed to determine what kind of effect on the environment when fine dust is generated. Fine dust (PM10) and ultrafine dust (PM2.5) showed a correlation of 0.76, confirming that a high correlation was formed, while PM2.5 and Co2 and O3 and Co2 had a strong correlation of 0.59. appeared to be PM2.5 and Co showed a correlation of 0.63, and Co2 and Co showed a strong correlation of 0.75. Finally, Mtemp, So2, and O3, which represent the average daily temperature, showed a weak correlation of 0.37. Therefore, when fine dust is generated, it can be seen that So2, Co, PM10, and PM2.5 appear together, and these substances adversely affect the environment.

In addition, the prediction model of ultrafine dust (HP2.5) showed somewhat under-prediction. However, in the case of the model using the selected input variable, the overall predicted concentration value was somewhat over-predicted. It was confirmed that the prediction performance of ultrafine dust (PH2.5) was improved through the use of variables correlated with fine dust (PH10). The performance of the fine dust prediction value, RMSE, and correlation value of the LSTM model used in this paper was high.

References

[1] S. Fuzzi, U. Baltensperger, K. Carslaw, S. Decesari, H. Denier vander Gon, M.C. Facchini, D. Fowler, I. Koren, B. Langford, U.Lohmann, E. Nemitz, S. Pandis, I. Riipinen, Y. Rudich, M. Schaap, J. G. Slowik, D. V. Spracklen, E.

Vignati, M. Wild, M. Williams, and S. Gilardoni, "Particulate matter, air quality and climate: Lessons learned and future needs," Atmospheric chemistry and physics, Vol. 15, No. 14, pp.8217-8299, 2015. DOI: https://doi.org/10.5194/acp-15-8217-2015

[2]] N. J. Hime, G. B. Marks, and C. T. Cowie, "A comparison of the health effects of ambient particulate matter air pollution from five emission sources," International Journal of Environmenta Research and Public Health, Vol. 15, No. 6, 2018. DOI: https://doi.org/10.3390/ijerph15061206

[3] Organization for Economic Co-operation and Development (OECD), The economic consequences of outdoor air pollution, Policy Highlights, 2016.

[4] H.-S. Choi, M. Kang, Y. C. Kim, and H. Choi, "Deep Learning-based Prediction of PM10 Fluctuation from Gwanak-gu Urban Area, Seoul, Korea," Journal of Soil and Groundwater Environment, vol. 25, no. 3, pp. 74–83, Sep. 2020. DOI: https://doi.org/10.7857/JSGE.2020.25.3.074

[5] T. Civas, F. F. Soulié, P. Gallinari, and S. Raudys, "Variable selection with neural networks," Neurocomputing, Volume 12.2-3, pp.223-248, 1996.

[6] R. May, G. Dandy, and H. Maier, "Review of input variable selection methods for artificial neural networks," *Artificial neural networks-methodological advances and biomedical applications*, 10:16004, pp.19-44, 2011.

[7] https://web.kma.go.kr/weather/forecast/timeseries.jsp

[8] https://www.airkorea.or.kr/index