Original Article
International Journal of Fuzzy Logic and Intelligent Systems
Vol. 19, No. 4, December 2019, pp. 242-249
http://doi.org/10.5391/IJFIS.2019.19.4.242

ISSN(Print) 1598-2645
ISSN(Online) 2093-744X

Developing a Big Data Analytic Model and a Platform for Particulate Matter Prediction: A Case Study

Sung-Hyun Kim[1], Dae-Sung Son[2], Min-Ho Park[3], and Hyun-Seok Hwang[4]
[1]National Information Society Agency (NIA), Daejeon, Korea
[2]OnKweather, Seoul, Korea
[3]KT, Seoul, Korea
[4]Hallym University, Chuncheon, Korea

Ijfis

Abstract

While we've made progress over the last century, qualities of air and drinking water are getting worse. For the health of the people, many countries recommend to wear a mask or not to go out in case of bad air quality. However, the lack of real-time measurement and poor prediction of the concentration of fine dust provide inaccurate alarm and information to residents. The real-time air quality measurement data is the big data of new field which should be complemented with the national meteorological data in order to meet the demands for the customers and public through the other relevant data. In this study we aim to build a system to predict PM (particulate matter)—microscopic dust generated by human activities, such as the burning of fossil fuels in vehicles, power plants and various industrial processes. We also develop an optimized path of sprinkler vehicles to improve air condition in a local city. The density of PM is affected by meteorological variables such as temperature, rainfall, humidity and wind speed. To predict the density of (fine) PM, we use machine learning techniques considering not only meteorological data from observatories, but spatial factors like pollutants, floating populations and traffic volumes of a certain location using LTE(Long Term Evolution) signals. We suggested a conceptual architecture of a system and illustrate a case study of preprocessing and analyzing the data gathering from a local province with results and discussions.

Keywords: Particulate matter, Big data, Prediction model, Artificial intelligence, Fine dust

Received: Sep. 30, 2019
Revised : Dec. 1, 2019
Accepted: Dec. 5, 2019

Correspondence to: Hyun-Seok Hwang
(hshwang@hallym.ac.kr)

1. Introduction

World Health Organization classified fine dust as the group 1 carcinogenic substance which causes a serious health risk along with direct impact on work productivity, academic performance and work efficiency. The inconvenience and the anxiety of the people due to the fine dust has been increasing, on the other hand the utilization/accumulation of the data on the air pollution regarding the fine dust and its countermeasures against the fine dust are not properly performed. The real-time measurement data about air quality is the big data of new field which should be complemented with the national meteorological data in order to meet the demands for public people and relevant industry.

The changes in the frequency of dust generation, changes in power usage patterns, outbreaks of the disease, and the movement pattern of people can be provided valuable information not only to the private sectors such as energy companies, electronics companies, medical institutions, and insurance companies, but also to the public sectors like local governments and schools. Air quality can be increased by dispatching and re-routing watering vehicles to appropriate areas on after predicting the future air pollution.

The study is to find out the correlations between the data and various activities of the people (such as health, economic activity, productivity, work efficiency) directly and indirectly by installing the air quality measurement which can be monitored for 24/7. With combining the air quality measurement data and various data in other fields can create the new big data with high value. This project stands for proposing a new value-added service through integration and analysis of data that relates to the fine dust and its relevant behaviors. To provide the public with a guidance/guidebook by establishing a big data platform of the fine dust, and to provide administrative guidance/guidebook to the public institutions and local governments and inform the integrated service of indoor and outdoor air management.

This study is organized as follows: Section 2 reviews related works of particular matter. Section 3 suggests a research framework regarding a big data analytic system and a model for predicting particular matter. In Section 4, we cover analytic results. We discuss the results and conclude our study with future research directions in Section 5.

2. Related Work

The literatures in predicting the density of particulate matter have several research directions. Main directions in these studies can be classified by prediction models and type of data used as shown in Table 1.

A machine learning approach is the most popular way of predicting the concentration of PM (particulate matter). Artificial neural network (ANN) is a common technique to represent features of forecasting the density of PM [1–10].

The second popular approach is a statistical analysis. Land use regression (LUR)–originally developed as a means to assess exposures from traffic-related air pollution–is a typical method to assess air pollution epidemiology [11–15].

Figures 1 and 2 show the number of annual papers published in relation to particulate matter and the word cloud used in the keywords, respectively.

Table 1. Main research directions in predicting the density of PM

Category	Description
Prediction model	
Machine learning	Artificial neural network, Support vector machine
Statistical	Land use regression
Ensemble	Random forest
Type of data	
Spatial data	Satellite data, Geographical
Pollutants data	Traffic
Meteorological data	Temperature, Rainfall, Humidity, Wind speed
Temporal data	Pollutants, Populations, Vehicles
Combined	Spatial + Pollutants + Meteorological + Temporal data

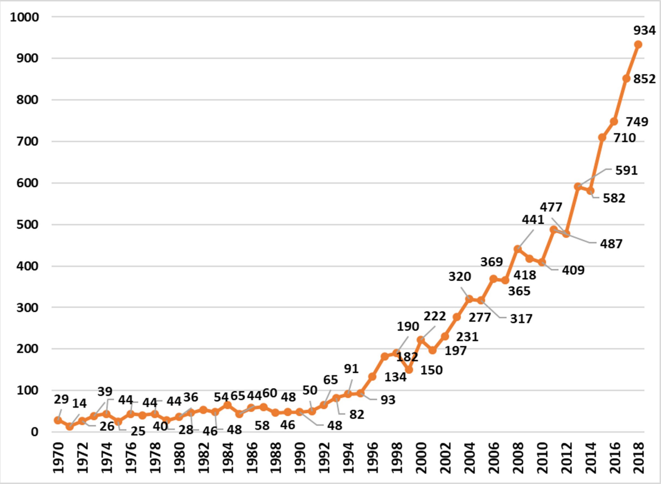


Figure 1. Number of published articles (1970–2018).

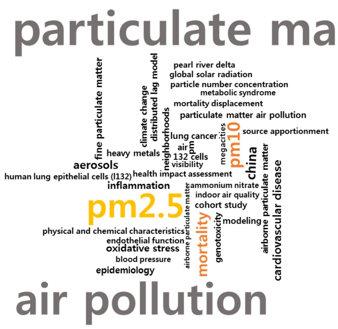


Figure 2. Word cloud of keywords in PM related articles.

3. A Research Method

We suggest the conceptual architecture of a big data analytic system for particulate matter as shown in Figure 3. To analyze the density of particulate matter in the air, it is required to gather data from various sources such as Korean government, Meteorological Agency, Internet of Things (IoT) sensors, human activities and pollutants. A lot of data have spatio-temporal characteristics–values with geological location and/or with time stamp. While some of data is stored in databases and transferred through FTP (file transfer protocol), others are collected by sensors and transmitted through LTE networks automatically.

Heterogeneous and large volume of data should be stored in huge repository and accessible through networks. We suggest the mash-up data is preprocessed and stored in cloud service for ubiquitous access to shared pools of configurable system resources. A pre-processing process can include data quality assurance, normalization, and standardization.

Correlation analysis is an inevitable process for variable reduction. Multiple competing machine learning models need to be used to enhance the prediction accuracy. The target variable is the density of PM10 or PM2.5 with a 1-hour time gap between input variables. Since the target values are valid only in device locations, we need to estimate the target values of shaded areas considering spatial characteristic of estimated locations such as distances from multiple adjacent devices, floating population, traffic volumes and number of pollutants. After building a feasible analytic results, we will disseminate the outcomes to the public through open API (application programming interface).

Currently, a limited number of devices and lack of prediction mechanism cause wide shaded areas. We considered two models to estimate the density of shaded areas: (i) naive model of predicting an average density value from nearest two devices' values, and (ii) interpolation model of a weighted average value from n-nearest devices' values. Weight will be the closeness between a shaded spot and an IoT device. Figure 4 shows an

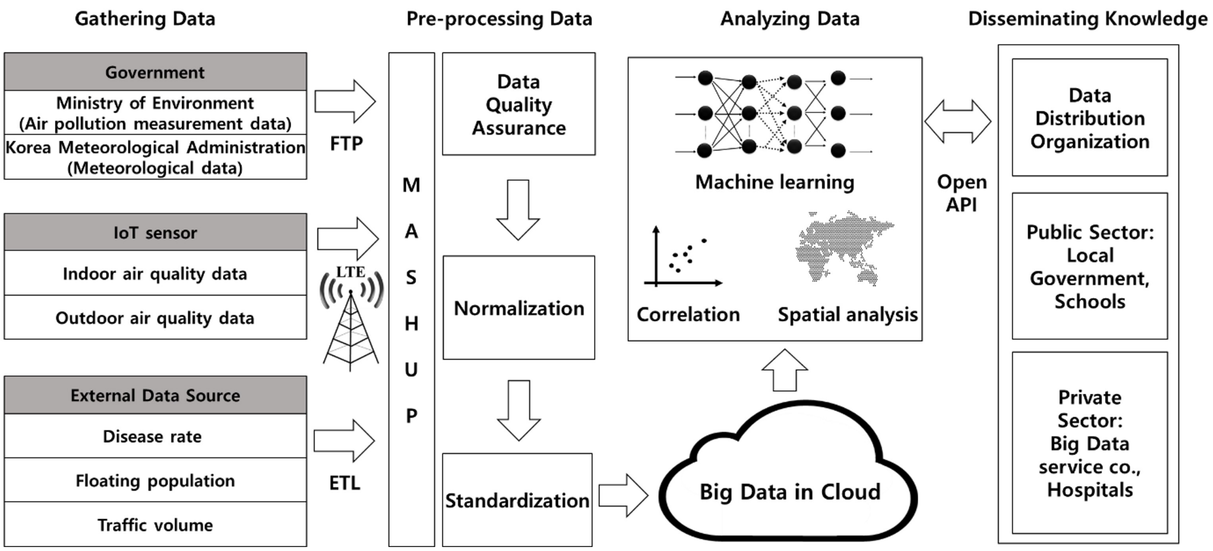
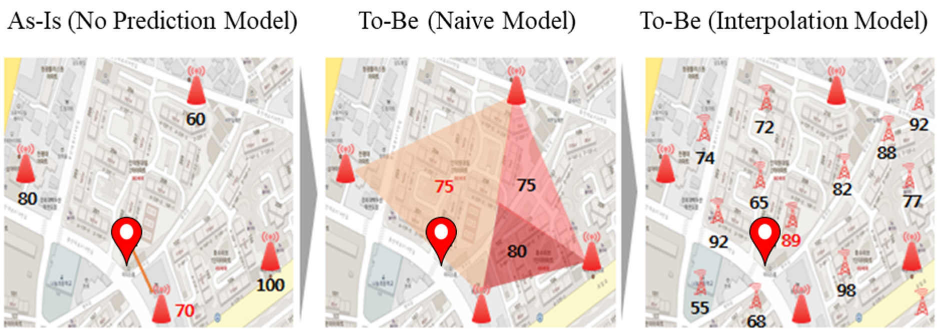


Figure 3. A conceptual architecture of big data analytic system for particulate matter (PM).


As-Is (No Prediction Model)

To-Be (Naive Model)

To-Be (Interpolation Model)

Figure 4. An As-Is model and To-Be models of predicting the densities of particulate matter in shaded areas.

As-Is model and To-Be models of predicting the densities of particulate matter in shaded areas.

$$V_t = \sum_i w_i V_i / \sum_i w_i, \tag{1}$$

where $w_i = 1/d_i$.

V_i denotes the measure value of spot i and d_i means the Euclidean distance between estimation spot and spot i. Eq. (1) means adjacent values will affect the predicted value in inverse proportion to the distance.

4. A Case Study

A case study is a nation-wide project to propose a new value-added service through integration and analysis of data that relates to particulate matter and its relevant behaviors. Fine dust measuring equipment is installed by the government at 320 branches nationwide and it measures the average value of the air quality every hour. In recent years, new IoT air quality measuring instruments have been introduced into the market, in combination with IoT technology.

Table 2. Specification of the big data system

Category	Detail
Operating system	CentOS release 6.9
Database	postgres (PostgreSQL) 8.4.20
Application language	
Batch Job	Shell script, crontab
Preprocessing	Stored procedure
Analysis language	Python 3.6.3
	Library: scikit learn v0.19.1

The aim of project is to provide the public with a guidance/guidebook by establishing a big data platform of the fine dust, and to provide administrative guidance/guidebook to the public institutions and local governments and inform the integrated service of indoor and outdoor air management. The specification of the big data system is shown in Table 2.

We chose Jeju Island as a project area. Jeju Island is the cleanest area in Korea and is isolated from other areas, making it the easiest area to control and observe pollutants. Table 3 shows the raw data and details for predicting particulate matter. In order to predict the value after 1 hour, we collected the input values measured hourly.

4.1 Gathering and Preprocessing Raw Data

We used the density of (fine) PM (PM2.5, PM10) as both an input variable and a target variable. The densities are measured by IoT devices linked to LTE network and is transmitted automatically. There is a 1-hour gap between the two variables. The two variables, way of measuring density and characteristics of device location, were excluded since they are not influential to the target variable.

The meteorological data–temperature, amount of rainfall, humidity, wind speed–were downloaded from the government's FTP server. If there is a missing value, it is replaced with another date value in the same time zone.

Data use the previous time value in case of missing values.

Human movements (floating population and estimated traffic volume) are measured by the number of people and vehicles connected to the mobile phone base stations. The numbers are

Table 3. Raw data for predicting particulate matter (PM)

Category	Details	Metric	Volume (instance)
Air quality measurement data	Time span: 2017/06/02-2017/09/25 Number of devices: 58 Measurement frequency: 1 min	PM10 PM2.5	439 MB (319 million cases)
Pollution source data	Time span: 2017/04/01-2017/09/30 Measurement frequency: 1 hr	Floating population Estimated traffic volume	1 GB (1.049 billion cases)
	Time span: 2017/09	Pollution volume Pollution grade Latitude-Longitude	76 kB (6,000 sources)
Meteorological data	Time span: 2017/06/02-2017/09/25 Measurement Frequency: 1 hr	Temperature Cumulative rainfall Humidity Wind speed	484 kB (112,000)

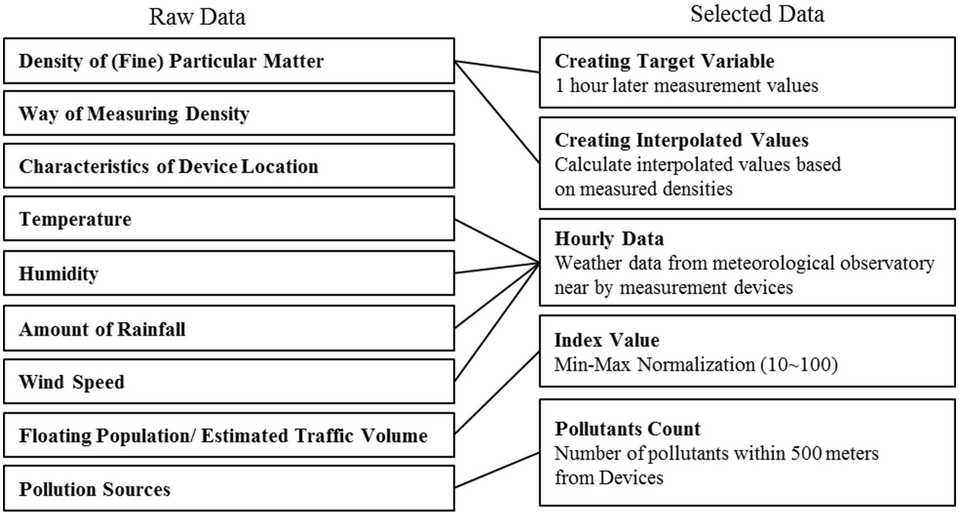

Raw Data
Selected Data
Density of (Fine) Particular Matter
Way of Measuring Density
Characteristics of Device Location
Temperature
Humidity
Amount of Rainfall
Wind Speed
Floating Population/ Estimated Traffic Volume
Pollution Sources
Creating Target Variable
1 hour later measurement values
Creating Interpolated Values
Calculate interpolated values based on measured densities
Hourly Data
Weather data from meteorological observatory near by measurement devices
Index Value
Min-Max Normalization (10~100)
Pollutants Count
Number of pollutants within 500 meters from Devices

Figure 5. Raw data mapping to input variables and target variable.

converted to 10–100 using min-max normalization. Pollution source is measured by the number of pollutants within 500 m from devices. Pollutants include manufacturing facilities, power plants, oil & gas refineries, and chemical factories.

Figure 5 shows the mapping between the raw data and the input & target variable(s).

5. Building Models and Predicting the Density

Figure 6 depicts the procedure of building models and predicting the density of particulate matter. 70% data was used for developing a prediction model and 30%, for performance test. Three competing models–neural network, multiple linear regression, and decision tree–compared their accuracies. The best model is used to predict 1-hour-later density of particulate matter.

The neural network model showed better R^2 and root-mean-squared error (RMSE) values than other models and was adopted as a fine dust prediction model. Although the performance difference is not large compared with other models, it is considered that the neural network is most suitable considering the improvement through continuous data accumulation. Table 4 shows the performance comparison results.

The density of target spot can be estimated by interpolation using Eq. (1). We gathered the density data from 58 IoT devices. However, we still need some interpolations to estimate the densities of shaded areas as shown in Figure 7. Considering accessibility and stability of density data, we use 5 adjacent measured densities to estimate the density of a shaded spot.

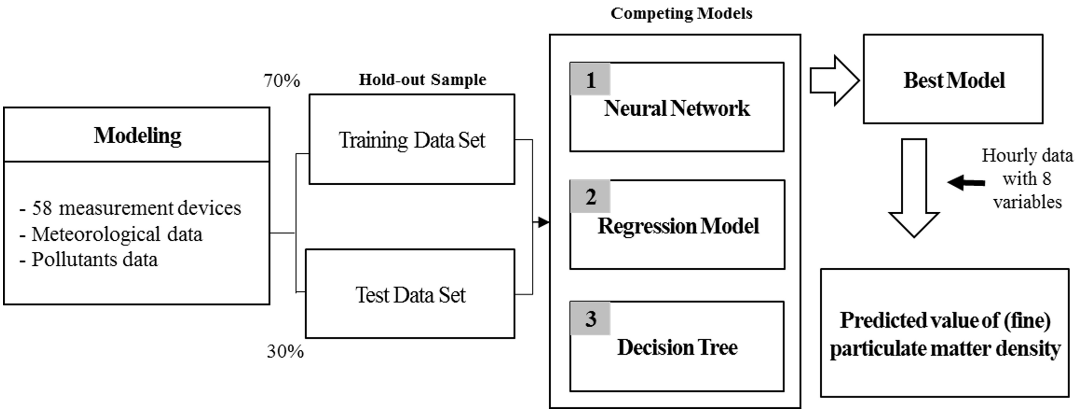


Figure 6. Procedure of building models and predicting the density.

Table 4. Prediction accuracies

Target variable	Model	Training set		Test set	
		R^2	RMSE	R^2	RMSE
PM10	Neural network	0.70823	9.90440	0.715764	9.89117
	Linear regression	0.70293	9.94345	0.712042	9.87822
	Decision tree	0.68144	10.28459	0.688197	10.28040
PM2.5	Neural network	0.71921	6.76133	0.71845	6.75422
	Linear regression	0.71337	6.79357	0.71265	6.76922
	Decision tree	0.71740	6.74638	0.71656	6.72331

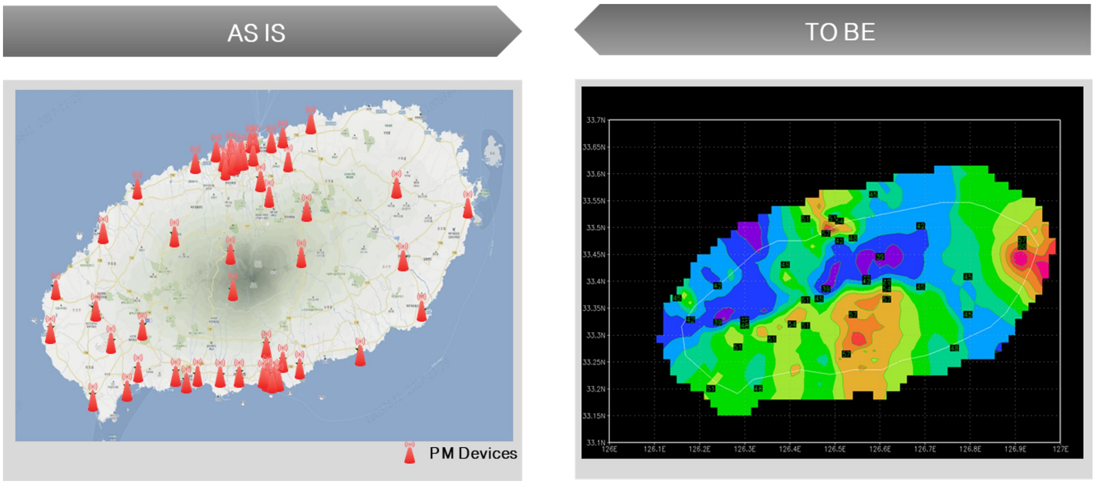


Figure 7. Location of IoT sensors and coverages.

6. Conclusion

We proposed an analytic system architecture for predicting PM density. Three data mining models are built and compared to find the best model for PM prediction. We also suggested an interpolation method to cover up the shaded areas. We expect that this study will be expanded to conduct a nationwide verification and to suggest some practical implications after further researches.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgements

This paper used some results from the research project funded by the Ministry of Science & ICT. This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (No. NRF-2018S1A3A2074932).

References

[1] M. D. Adams and P. S. Kanaroglou, “Mapping real-time air pollution health risk for environmental management: combining mobile and stationary air pollution monitoring with neural network models,” *Journal of Environmental Management*, vol. 168, pp. 133-141, 2016. https://doi.org/10.1016/j.jenvman.2015.12.012

[2] G. Grivas and A. Chaloulakou, “Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece,” *Atmospheric Environment*, vol. 40, no. 7, pp. 1216-1229, 2006. http://dx.doi.org/10.1016/j.atmosenv.2005.10.036

[3] H. He, W. Z. Lu, and Y. Xue, “Prediction of particulate matter at street level using artificial neural networks cou-

pling with chaotic particle swarm optimization algorithm," *Building and Environment*, vol. 78, pp. 111-117, 2014. http://dx.doi.org/10.1016/j.buildenv.2014.04.011

[4] J. Hooyberghs, C. Mensink, G. Dumont, F. Fierens, and O. Brasseur, "A neural network forecast for daily average PM10 concentrations in Belgium," *Atmospheric Environment*, vol. 39, no. 18, pp. 3279-3289, 2005. http://dx.doi.org/10.1016/j.buildenv.2014.04.011

[5] X. Leng, J. Wang, H. Ji, H., Q. Wang, H. Li, X. Qian, F. Li, and M. Yang, "Prediction of size-fractionated airborne particle-bound metals using MLR, BP-ANN and SVM analyses," *Chemosphere*, vol. 180, pp. 513-522, 2017. https://doi.org/10.1016/j.chemosphere.2017.04.015

[6] J. B. Ordieres, E. P. Vergara, R. S. Capuz, and R. E. Salazar, "Neural network prediction model for fine particulate matter (PM2.5) on the US–Mexico border in El Paso (Texas) and Ciudad Ju?rez (Chihuahua)," *Environmental Modelling & Software*, vol. 20, no. 5, pp. 547-559, 2005. http://dx.doi.org/10.1016/j.envsoft.2004.03.010

[7] P. Perez and J. Reyes, "Prediction of maximum of 24-h average of PM10 concentrations 30 h in advance in Santiago, Chile," *Atmospheric Environment*, vol. 36, no. 28, pp. 4555-4561, 2002. http://dx.doi.org/10.1016/S1352-2310(02)00419-3

[8] U. A. Sahin, C. Bayat, and O. N. Ucan, "Application of cellular neural network (CNN) to the prediction of missing air pollutant data," *Atmospheric Research*, vol. 101, no. 1-2, pp. 314-326, 2011. http://dx.doi.org/10.1016/j.atmosres.2011.03.005

[9] A. Suleiman, M. R. Tight, and A. D. Quinn, "Assessment and prediction of the impact of road transport on ambient concentrations of particulate matter PM10," *Transportation Research Part D: Transport and Environment*, vol. 49, pp. 301-312, 2016. http://dx.doi.org/10.1016/j.trd.2016.10.010

[10] Y. Zhan, Y. Luo, X. Deng, H. Chen, M. L. Grieneisen, X. Shen, L. Zhu, and M. Zhang, "Spatiotemporal prediction of continuous daily PM2.5 concentrations across China using a spatially explicit machine learning algorithm," *Atmospheric Environment*, vol. 155, pp. 129-139, 2017. https://doi.org/10.1016/j.atmosenv.2017.02.023

[11] C. Ho, C. Chan, C. Cho, H. Lin, J. Lee, and C. Wu, "Land use regression modeling with vertical distribution measurements for fine particulate matter and elements in an urban area," *Atmospheric Environment*, vol. 104, pp. 256-263, 2015. https://doi.org/10.1016/j.atmosenv.2015.01.024

[12] M. Hochadel, J. Heinrich, U. Gehring, V. Morgenstern, T. Kuhlbusch, E. Link, H. Wichmann, and U. Kramer, "Predicting long-term average concentrations of traffic-related air pollutants using GIS-based information," *Atmospheric Environment*, vol. 40, no. 3, pp. 542-553, 2006. https://doi.org/10.1016/j.atmosenv.2005.09.067

[13] X. Meng, Q. Fu, Z. Ma, L. Chen, B. Zou, Y. Zhang, et al., "Estimating ground-level PM10 in a Chinese city by combining satellite data, meteorological information and a land use regression model," *Environmental Pollution*, vol. 208, pp. 177-184, 2016. https://doi.org/10.1016/j.envpol.2015.09.042

[14] H. Merbitz, S. Fritz, and C. Schneider, "Mobile measurements and regression modeling of the spatial particulate matter variability in an urban area," *Science of the Total Environment*, vol. 438, pp. 389-403, 2012. http://dx.doi.org/10.1016/j.scitotenv.2012.08.049

[15] G. Pereira, H. J. Lee, M. Bell, A. Regan, E. Malacova, B. Mullins, and L. D. Knibbs, "Development of a model for particulate matter pollution in Australia with implications for other satellite-based models," *Environmental Research*, vol. 159, pp. 9-15, 2017. http://dx.doi.org/10.1016/j.envres.2017.07.044

Sung-Hyun Kim is director of K-ICT Big Data Center of National Information Society Agency. He has been working on performance management and big data. He holds a Ph.D. from Sungkyunkwan University Business School in management information systems, a MBA from Korea University. Before joining NIA, he worked at Samsung SDS as a Senior Consultant. His interests are cloud big data and data business model.
E-mail: kimcon@nia.or.kr

Dae-Sung Son is a director of ICT Business Department at Kweather Co. Ltd., in Korea. He received a master's degree from the Graduate School of Industrial Technology at Kookmin University. He leads the development of air quality monitoring system based on IOT technology and the analysis of a measured data.
E-mail: son@onkweather.com

Min-Ho Park is a Team Leader of Environment Platform Business TF in Korea Telecom Corp., South Korea. He received his master's degree in Electronic Engineering from Yonsei University, Korea. He is a member of environment advisory committee of Seoul Metropolitan City, Gyeonggi and Seoul Province in Korea.
E-mail: min-ho.park@kt.com

Hyun-Seok Hwang is a Professor of Business Administration and a research fellow of Hallym Business Research Institute at Hallym University, Chuncheon, Korea. He received his Ph.D. in Industrial Engineering from the Pohang University of Science and Technology (POSTECH), Korea. His current research focuses on big data analytics, opinion mining, and interaction design.
E-mail: hshwang@hallym.ac.kr