Journal of the Korean Association of Geographic Information Studies, 24(1) 2021, pp.92-111 ISSN 1226-9719(Print)
https://doi.org/10.11108/kagis.2021.24.1.092 ISSN 2287-6952(Online)

기상 데이터와 미세먼지 데이터를 활용한 머신러닝 기반 미세먼지 예측 모형*

김혜림[1] · 문태헌[2※]

Machine learning-based Fine Dust Prediction Model using Meteorological data and Fine Dust data*

Hye-Lim KIM[1] · Tae-Heon MOON[2※]

요 약

미세먼지는 질병, 산업 · 경제에 부정적인 영향을 미치고 있어 국민들은 미세먼지에 대해 예민하게 반응하고 있다. 따라서 미세먼지의 발생을 예측할 수 있다면, 미리 대응책을 마련할 수 있어 생활과 경제에 도움이 될 수 있다. 미세먼지의 발생은 기상과 미세먼지 배출원의 밀집 정도에 영향을 받는다. 산업부문은 미세먼지 배출량이 가장 많으며, 그 중에 산단은 공장들이 미세먼지 배출원이 되어 더 많은 미세먼지를 배출하는 문제가 있다. 본 연구는 지방도시에서 노후산업단지가 있는 지역을 선정하여, 미세먼지를 일으키는 요인을 탐색하고, 미세먼지 발생을 예측할 수 있는 예측모형을 개발하고자 한다. 기상 데이터와 미세먼지 관련 데이터를 활용하였고, 다중회귀분석을 통해 미세먼지 발생에 영향을 미치는 변수를 추출하였다. 이를 토대로 머신러닝 회귀학습기 모형으로 학습하여 예측력이 높은 모형을 추출하였고, 검증용 데이터를 이용하여 예측 모형의 성능을 검증하였다. 그 결과, 예측력이 높은 모형은 선형회귀모형, 가우스 과정 회귀모형, 서포트 벡터 머신으로 나타났으며, 훈련용 데이터의 비율과 예측력은 비례하지 않은 것으로 나타났다. 또한 예측치와 실측치 차이의 평균치는 크지 않지만, 미세먼지 실측치가 높을 때, 예측력이 다소 떨어지는 것으로 나타났다. 본 연구의 결과는 지자체 데이터 허브를 통해 기상데이터와 관련 도시 빅데이터를 결합함으로써 보다 체계적이고 정밀한 미세먼지 예측 서비스로 개발이 가능할 것이며, 스마트산단의 발전을 촉진하는 계기가 될 것이다.

주요어 : 미세먼지, 기상, 빅데이터, 머신러닝, 스마트산단

2021년 02월 22일 접수 Received on February 22, 2021 / 2021년 03월 10일 수정 Revised on March 10, 2021 / 2021년 03월 10일 심사완료 Accepted on March 10, 2021

* 이 제작물은 지자체-대학 협력기반 지역혁신 사업비에서 지원하여 제작됨

1 경상대학교 도시공학과 대학원생 Master Student, Dept. of Urban Engineering, Gyeongsang National University
2 경상대학교 도시공학과 교수 Professor, Dept. of Urban Engineering, ERI. Gyeongsang National University
※ Corresponding Author E-mail: thmoon@gnu.ac.kr

ABSTRACT

As fine dust negatively affects disease, industry and economy, the people are sensitive to fine dust. Therefore, if the occurrence of fine dust can be predicted, countermeasures can be prepared in advance, which can be helpful for life and economy. Fine dust is affected by the weather and the degree of concentration of fine dust emission sources. The industrial sector has the largest amount of fine dust emissions, and in industrial complexes, factories emit a lot of fine dust as fine dust emission sources. This study targets regions with old industrial complexes in local cities. The purpose of this study is to explore the factors that cause fine dust and develop a predictive model that can predict the occurrence of fine dust. weather data and fine dust data were used, and variables that influence the generation of fine dust were extracted through multiple regression analysis. Based on the results of multiple regression analysis, a model with high predictive power was extracted by learning with a machine learning regression learner model. The performance of the model was confirmed using test data. As a result, the models with high predictive power were linear regression model, Gaussian process regression model, and support vector machine. The proportion of training data and predictive power were not proportional. In addition, the average value of the difference between the predicted value and the measured value was not large, but when the measured value was high, the predictive power was decreased. The results of this study can be developed as a more systematic and precise fine dust prediction service by combining meteorological data and urban big data through local government data hubs. Lastly, it will be an opportunity to promote the development of smart industrial complexes.

KEYWORDS : Fine Dust, Weather, Big Data Machine learning,Smart Industrial Complex

서 론

미세먼지는 호흡기계, 심혈관계 질병의 발생 및 악화와 같이 인간의 건강뿐만 아니라 항공기, 선박 운영제한 등 사회 · 경제적으로도 좋지 않은 영향을 미친다. Kyung *et al.*(2015)에 의하면, 초미세먼지(PM2.5) 농도가 10㎍/㎥ 증가할수록 폐암 발생이 9% 증가하고, 미세먼지(PM10) 농도가 10㎍/㎥ 증가할수록 만성폐쇄성폐질환 관련 입원이 2.7%, 만성폐질환 관련 사망이 1.1% 증가한다고 한다. 이와 같이 미세먼지는 국민건강을 위협할 뿐 아니라 초정밀 고순도를 유지해야 하는 반도체와 같은 첨단산업의 공정과 생산, 먼지가 존재하는 환경에서 생산할 수 없는 주사제의 생산에도 영향을 주는 것으로 알려져 있다. 또한 국민들은 미세먼지에 대한 경각심이 증대되면서 야외활동을 기피하고, 이는 지역 경제에도 영향을 미치게 된다. 미세먼지 발생이 심각한 날에는 가시거리가 확보되지 않아 교통수단이 결항, 지연, 취소되기도 한다.

이와 같이 미세먼지의 발생이 사회 · 경제적으로 다양한 분야에 부정적인 영향을 미치면서, 미세먼지를 뜻하는 Dust와 공포증을 뜻하는 Phobia, 두 단어로 합성어를 만들어 '더스트포비아' 라는 신조어가 생겨날 만큼, 국민들의 미세먼지에 대한 관심과 두려움은 크다. 따라서 미세먼지 발생을 미리 예측하여 시민과 산업현장에 알려 미리 대비할 수 있도록 한다면 시민건강을 증진시키고, 산업활동이나 경제활동에 많은 도움을 줄 것이다.

우리나라는 2004년 4월부터 전국의 대기오염측정망에서 측정되는 미세먼지를 포함한 대기오염도 자료를 수집·관리하는 국가대기오염정보관리시스템(NAMIS)을 구축하여 정보를 제공하고 있다. 이후 2005년 12월 28일에는 '에어코리아' 라는 전국 실시간 대기오염도 공개 홈페이지를 구축하여 전 국민이 대기오염도 자료를 쉽고 편리하게 접할 수 있도록 하였다. 이는 실시간 측정과 실시간 알림이라는 점에서 많은 발전이 있기는 하지만 미리 대비할 수 있는 시간적 여유가 없고, 사후 대책만 가능한 한계점이 있다. 따라서 미세먼지 발생을 예측할 수 있는 방법의 개발이 필요하다. 특히 4차 산업혁명시대에 어울리는 스마트기술과 첨단기법을 활용하여 보다 정밀하고 예측력이 좋은 방법의 개발이 필요하다.

미세먼지의 발생과 주성분은 기상과 매우 관련이 높으며, 그 지역에 미세먼지 배출원이 얼마나 존재하는지 그리고 배출량은 어느 정도인가에 따라 크게 좌우된다. 국내 미세먼지 배출량을 분야별로 보면 산업 37%, 수송 28%, 생활 20%, 발전 15% 순으로 높아, 산업계에서 미세먼지 저감과 관리 대책이 특히 강조된다(Ministry of Environment, 2017). 특히 도시내에 입지한 산업단지가 주요 배출원인 것으로 알려져 있다. 산단은 화석연료 기반의 에너지 공급구조로 다량의 온실가스가 배출되며, 산단이 산업부문 온실가스배출량의 76.8%를 차지할 정도로 위협적인 요소이다(Ministry of Trade, Industry and Energy, 2020).

지역적으로 보면, 다양한 미세먼지 배출 시설이 있는 대도시와는 달리, 노후산업단지(이하 노후산단)가 존재하고 시가지는 크지 않은 지방 중소도시의 경우는 미세먼지에 더욱 크게 영향을 받을 수 있으며, 시민들에게 일상화되어 감각이 무뎌진 대도시보다 더 민감하게 반응할 수 있다. 따라서 본 연구에서는 지방도시를 대상으로 하되, 미세먼지와 환경에 취약한 노후산단이 있는 지역을 선정하여, 미세먼지를 일으키는 요인을 탐색하고, 미세먼지 발생을 예측할 수 있는 예측모델을 개발하고자 한다.

노후산업단지 개조와 스마트산업단지 사업 개요

한국산업단지공단 전국산업단지현황통계(2020)에 의하면, 우리나라의 산업단지는 2020년 3분기를 기준으로 전국에 총 1,255개가 지정되어 있으며, 산업단지는 지난 50년간 우리나라의 경제성장을 이끌어왔다. 하지만 산업단지가 노후화되면서 공장 가동률 저하, 생산성 감소는 물론 환경오염 등의 문제가 대두되고 있다. 이에 정부에서는 2019년 11월에 발표한 산업단지 대개조 계획에 따라 산업단지 대개조 사업을 추진하여 제조혁신, 쾌적한 근로 · 정주환경, 창업과 신산업 활성화를 도모하고자 하였지만, 이 사업은 산단의 물리적인 환경개선을 중심으로 추진되어 근본적 개조에는 한계가 있다.

이와 유사하게 정부는 산업단지를 재생하려는 정책을 마련하여 전국적으로 확대해 나가려고 한다. 산단재생사업은 노후산단와 그 주변지역을 도시재생활성화 지역으로 지정하여 민간의 참여와 지역 중심으로 사업추진을 위한 기반을 강화하려고 하지만, 현장에서는 복잡한 이해관계와 높은 지가 등의 문제로 그다지 원만하게 진행되지 못하고 있는 실정이다. 게다가 물리적 개선을 위주로 하는 사업이라 노후산단을 산업구조 고도화시키고, 미래형 산업단지로 변모시키는 데는 많은 한계가 있다.

한편, 정부에서는 산업단지의 제조혁신을 위해 스마트공장을 보급할 뿐 아니라, 산단에 디지털 인프라 구축이 미흡하기 때문에 산단내 기업의 산업혁신을 유도하고자 산업단지 전체를 스마트화 하려는 시도도 있다. '디지털' 과 '그린' 이 융합된 미래형 혁신 스마트그린산업단지로 개념을 설정하고 있으며, 2020년 7월에 발표된 '한국판 뉴딜 종합계획' 의 10대 과제 중 하나로 선정할 정도로 중요한 사업으로 추진되고 있다. 그 해 9월에는 스마트그린산단 실행전략을 발표하면서 전국에 7개 산단을 지정하여 글로벌 경쟁력을 갖춘 친환경 첨단산업거점으로 전환하려는 계획을 가지고 있다.

이상과 같이 최근 산단과 관련하여 중요한 사

업이 진행되고 있으나 본 연구에서는 세계적으로 이미 4차 산업혁명시대에 진입하면서 스마트화에 관심이 집중되고 있으므로 산단의 스마트화에 초점을 두고자 한다. 스마트산단에 관해서는 앞서 언급한 바와 같이 정부에서 스마트그린산단의 개념을 정립하고 추진하고 있으므로 여기에서 필요한 부분을 검토하여 연구의 범위를 정하고자 한다.

정부의 스마트그린산단 정책을 보면 3가지 이슈별 전략을 설정하고 있는데, 산업 측면에서 글로벌 선도 첨단산단으로 전환, 공간 측면에서 저탄소 친환경 산단으로 혁신을, 사람 측면에서는 청년희망키움 공간으로 탈바꿈과 같이 정리하고 있다. 하지만 본 연구에서는 산단 자체의 스마트화를 대상으로 하고자 하며, 그 중에서 친환경산단에 중점을 두고자 한다. 본 연구는 노후 산업단지를 스마트 산업단지로 재생 및 발전시키기 위한 과정 중의 하나인 환경 분야에서 스마트 기술을 적용해 미세먼지 발생을 예측하는 방법을 개발해 보고자 한다.

연구 대상지 선정 및 현황

연구 대상지는 경남 진주와 상평일반산업단지로 하였다. 상평산단은 1978년 3월 15일에 지정되어 기계, 운송장비 관련 제조 업종을 주력으로 운영되고 있다. 2020년 현재 총 599개의 공장이 입주해있으며, 입주업종별 분포도는 그림 1과 같다. 이 중 351개 공장이 제조업 공장으로, 전체 공장의 58.6%를 차지한다.

상평산단의 공장별 대기오염물질 발생량 분포는 그림 2와 같고, 1종 사업장(연간 80톤 이상 배출하는 사업장) 2곳, 2종 사업장(연간 20톤 이상 80톤 미만 배출하는 사업장) 7곳, 3종 사업장(연간 10톤에서 20톤 미만 배출하는 사업장) 3곳, 4종 사업장(연간 2톤 이상 10톤 미만 배출하는 사업장) 11곳, 5종 사업장(연간 2톤 미만 배출하는 사업장) 32곳이다. 그림 1과 그림 2를 비교해 보면, 대기오염물질 발생량이 1

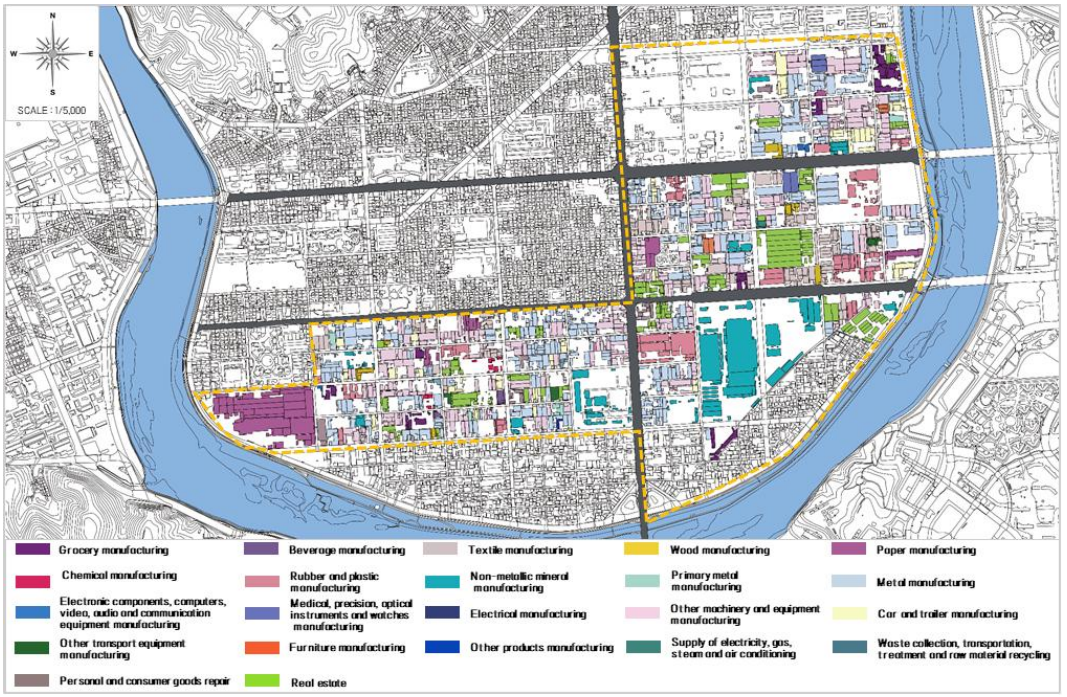

FIGURE 1. Sangpyeong industrial complex industry map

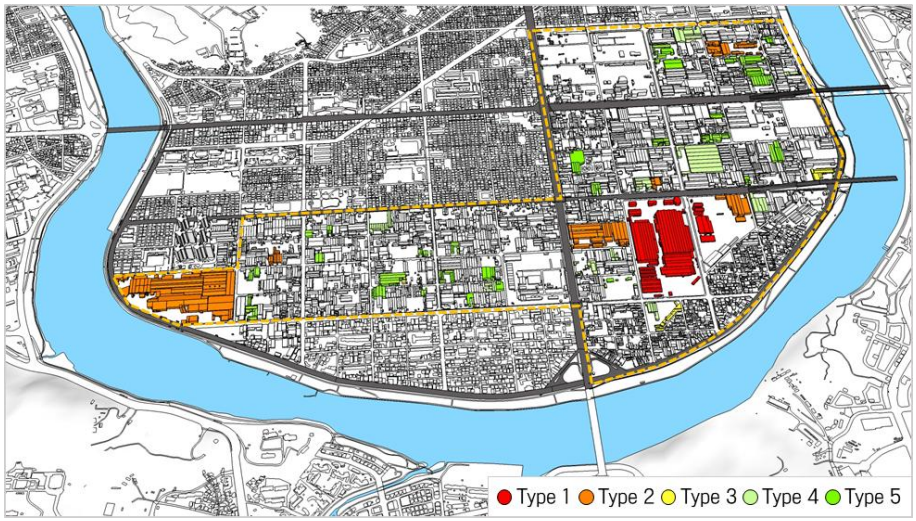

FIGURE 2. Distribution of factories generating air pollution in Sangpyeong industrial complex

종으로 분류된 공장의 업종은 전기, 가스, 증기 및 공기 조절 공급업, 2종은 전기, 가스, 증기 및 공기 조절 공급업과 금속 가공제품 제조업, 섬유제품 제조업, 고무 및 플라스틱 제조업으로 나타났다. 3종은 섬유제품 제조업과 식료품 제조업, 펄프, 종이 및 종이제품 제조업이며, 4종은 자동차 및 트레일러 제조업, 전기, 가스, 증기 및 공기 조절 공급업, 금속 가공제품 제조업, 1차 금속 제조업, 기타 기계 및 장비 제조업 등으로 나타났다. 5종으로 분류된 공장의 업종은 식료품 제조업, 기타 기계 및 장비 제조업, 섬유제품 제조업, 펄프, 종이 및 종이제품 제조업, 비금속 광물제품 제조업, 금속 가공제품 제조업, 자동차 및 트레일러 제조업, 화학 물질 및 화학제품 제조업, 고무 및 플라스틱제품 제조업, 개인 및 소비용품 수리점, 기타 운송장비 제조업 등이다.

상평산단에도 노후화가 진행되는 과정에서 스마트산단으로 변신하기 위해 부분적으로 노력하고 있다. 2020년 국가 인프라 지능정보화 사업 대상지로 선정되어, 산단의 인프라를 지능화하는 사업이 진행되고 있으며, 총 30개의 공기질 센서와 4개의 미세먼지 전광판이 설치되었다. 그 위치는 그림 3과 같으며, 상평산단에 설치된 공기질 센서로부터 PM2.5, PM10, NO_2, O_3, CO, H_2S, NH_3, SO_2, TVOC 농도가 실시간으로 수집되고 있다.

본 연구는 상평산단에서 수집되는 대기오염물질 관련 빅데이터를 활용하여, 미세먼지 농도의 증감에 영향을 주는 요인을 찾고, 이를 통해 산단의 미세먼지 발생량을 실시간으로 예측해 보고자 한다. 데이터는 에어코리아(https://www.airkorea.or.kr/)에서 제공하는 미세먼지 관련 빅데이터와 기상자료개방포털(https://data.kma.go.kr/)에서 제공하는 기상 관련 빅데이터를 활용하고자 한다.

문헌연구

미세먼지 발생량 예측과 관련한 연구는 다수 진행되고 있으나, 연구 대상지나 분석방법, 예측변수 등에서 다양하다. 연구 대상지는 하나의 시를 대상지로 하는 경우가 많고, 지역별로는 수도권 또는 서울시를 연구한 사례가 대부분이다(Lee and Lee, 2020; Kim, 2020; Cha and Kim, 2018). 이외에도 우리나라 6개 대도시 지역을 대상으로 한 연구(Jeon and Son, 2018),

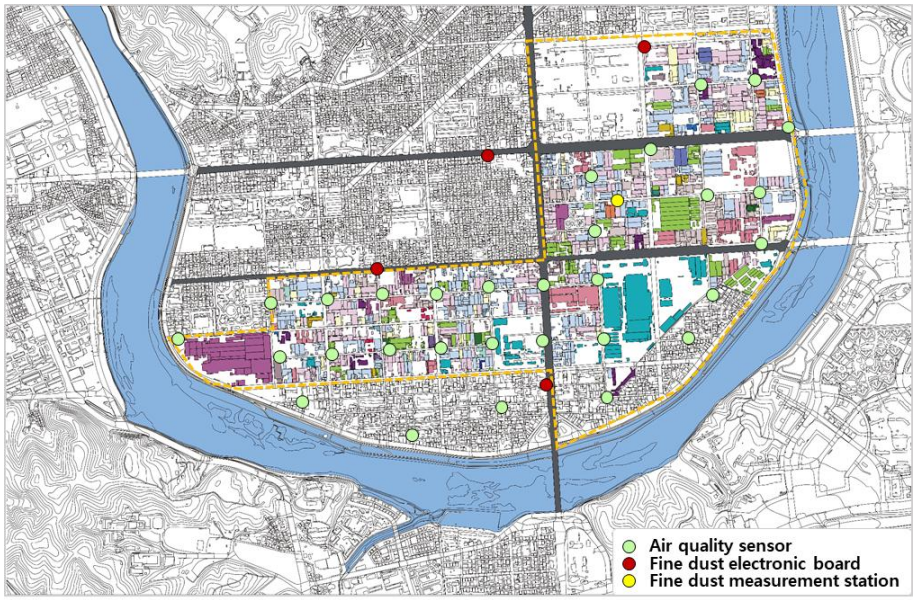

FIGURE 3. Location of sensors for air quality and fine dust in Sangpyeong industrial complex

실내 공간을 대상으로 한 연구(Yeo and Kim, 2019), 도시철도 공간을 대상으로 한 연구(Yoon *et al.*, 2018) 등이 소수 있다. 하지만 지방도시를 대상으로 한 연구사례는 발견하기 어렵다.

분석 방법은 최근 빅데이터 수집이 용이해지고, 머신러닝(machine learning) 기법이 발전함에 따라, 미세먼지와 같은 대기 환경을 예측하는 분야에서도 새로운 분석 방법이 도입되고 있다. 머신러닝은 인공지능의 한 분야로서, 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야로서, 알고리즘을 이용해 데이터를 분석하고, 분석을 통해 학습하며, 학습한 내용을 기반으로 판단이나 예측을 한다(Lee and Moon, 2017). 머신러닝은 본 연구와 같이 미세먼지 예측 외에도, 부동산 가격지수 예측(Bae and Yu, 2018), 범죄발생 위험지역 예측(Heo *et al.*, 2018), 공장 에너지 사용량 분석(Sung and Cho, 2019) 등 다양한 분야에서 활용되고 있다.

하지만 최근의 미세먼지 예측 관련 연구는 머신러닝에서 대부분 하나 또는 두세개 정도로 소수의 모형을 한정하여 연구를 진행하고 있다(Lee and Lee, 2020; Kim, 2020; Cha and Kim, 2018). 또한 시계열 분석도 활용하고 있으며, 여기서는 LSTM 모형이 많이 활용되고 있다(Kim *et al.*, 2019; Cho *et al.*, 2019; Lee and Lee, 2020). 본 연구에서는 예측 모형을 몇 가지로 한정하는 것이 아니라 여러 가지 회귀학습기 모형을 이용하여 최적 예측 모형을 찾고자 한다.

미세먼지 예측에 활용하는 예측변수 데이터로는 주로 기상 관련 데이터와 대기오염 물질 및 미세먼지 관련 데이터가 활용되고 있으며(Lee and Lee, 2020; Kim, 2020; Lim, 2019), 중국의 미세먼지 농도 데이터를 추가적으로 활용하기도 한다(Jeon and Son, 2018). 또한 최근 주목받고 있는 무인항공기, 드론에 대기측정센서를 부착하여 실시간으로 수집되는 대기 데이터를 활용하기도 하며(Kim *et al.*, 2019), 인공위성 데이터를 활용하기도 한다(Lee and Jeong, 2019).

연구 사례를 종합해 보면, 대상지는 주로 수

TABLE 1. Comparison of research cases

Case	Predictor	Area	Date	Analysis	predict level	predict power evalustion
Lee and Lee (2020)	Weather data, Air pollutant data, Fine dust data	Seoul	2015.1.1.– 2018.12.31	Random forest	Classification (daily)	F1
Kim (2020)	Weather data, Air pollutant data,	Seoul	2016.1.1.– 2019.6.30	XGBoost, Ensemble	Classification (Hourly)	Accuracy , Specificity, Responsiveness, Precision, F1
Jeon and Son (2018)	Weather data, Air pollutant data, Fine dust in China, Season	Seoul Gangnam, Busan Haeundae, Incheon Bupyeong, Daejeon Seo-gu, Gwangju Buk-gu, Daegu Dalseo	2010.1.1.– 2015.12.31	Deep neural network model	Classification (daily)	Accuracy
Cha and Kim (2018)	Weather data, air pollutant data	Seoul	2014.1.1.– 2017.6.31	Artificial neural network, K-Nearest Neighbor	numerical (daily)	Accuracy, Error rate
Kim *et al.* (2019)	Drone air measurement sensor data	–	21,600(s)	LSTM(Long Short-Term Memory)	numerical (20초)	RMSE
Lim (2019)	Weather data, Air pollutant data, Date data	Seoul	2014.1.1.– 2017.9.30	SVM, ANN	Classification (Hourly)	Match degree, Precision, Responsiveness
Oh (2017)	Fine dust data	Seoul	2001-2015	n□ive , Regression analysis, ARIMA, Exponential smoothing, Time series analysis	numerical (Monthly average)	Coefficient of determination, RMSE
Yoon *et al.* (2018)	Rail Wear Measurements	Urban Railway (all sections of Line S)	2009 (2.15-3.5), 2010 (8.26-9.18)	Create calculation formula	numerical (For 1 train operation)	–
Yeo and Kim (2019)	Children activities	Daycare indoor	Spring, Fall (2 times in 5 days)	Lumped model	numerical (By activity)	Linear comparison graph of predicted and measured values
Sohn and Kim (2015)	Weather data, China GTS data, Air pollution materials data	Seoul	2012.5.1.– 2013.8.12., 2014.1.1.– 2014.7.31	Multiple regression model, Threshold Regression Model	numerical (daily)	Correlation coefficient, RMSE

도권이나 서울을 중심으로 진행되고 있으며, 분석방법은 시계열 분석이나 머신러닝 기법이 활용되고 있으나 주로 몇 가지 모형으로 한정하여 진행되고 있다. 예측변수로는 주로 기상 관련 데이터와 대기오염 물질 및 미세먼지 관련 데이터를 활용하고 있으나 무인항공기나 드론과 같이 새로운 방법을 시도하는 사례도 있다.

이에 본 연구의 차별성으로는 지방 중소도시 중 노후산단을 대상으로 진행하였으며, 방법론적으로 기상 및 미세먼지 빅데이터를 활용하여 다양한 머신러닝 기법 기반의 미세먼지 일 발생농도 예측모형을 개발하는 것이다.

연구방법

연구방법은 그림 4와 같다. 기상자료 개방 포털에서 수집한 기상 자료와 에어코리아에서 수집한 미세먼지 자료를 기초로 데이터 형식을 통일하였다. 다음, 날짜 및 시간 자료, 기상 자료, 미세먼지 자료로 구성된 하나의 데이터셋을 구

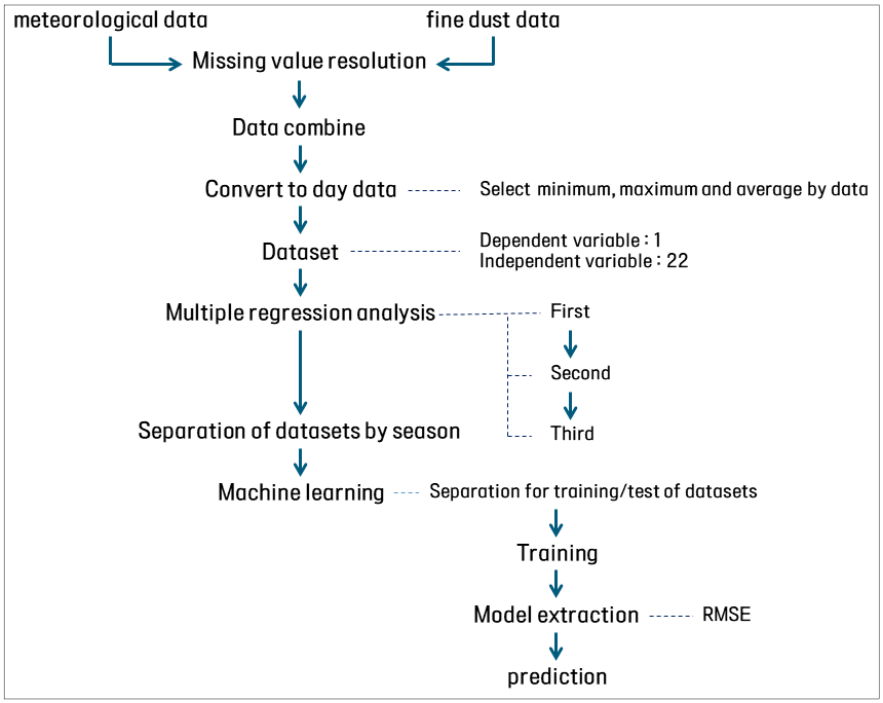

FIGURE 4. Research flow

축하였다. 이는 시간 자료이어서, 본 연구에서는 일 자료로 변환하여 사용하였다. 그 과정에서 각 변수들을 최소값, 최대값, 평균값 등 어떤 형태로 활용할지 결정하여 최종 데이터셋을 구축하였다. 최종 데이터셋은 종속변수 1개와 독립변수 22개로 구성되며, 총 731일간이다.

22개의 독립변수 중 종속변수 PM10에 유의한 영향을 주는 변수를 찾기 위해 다중회귀분석을 실시하였다. 1차 다중회귀분석에서 유의수준 5%를 기준으로 유의하지 않은 독립변수 13개를 제거하였으며, 2차 다중회귀분석을 실시하여 1차 다중회귀분석 후 남은 9개의 독립변수 중 독립변수 간에 강한 상관관계가 있는 2개의 독립변수를 다시 제거하였다. 최종적으로 3차 다중회귀분석을 실시하여 2차 다중회귀분석 후 남은 7개의 독립변수로 최종 회귀식을 도출하였다.

머신러닝에 앞서, 다중회귀분석 후 7개의 독립변수와 1개의 종속변수로 구성된 데이터셋을 계절별로 분리하여, 70:30, 80:20, 90:10 세 가지 비율로 랜덤하게 2회씩 분류하여 총 24개의 훈련용 데이터와 24개의 검증용 데이터를 구축하였다. 이 중 훈련용 데이터 각각을 19개의 머신러닝 회귀학습기 모형으로 학습하여 RMSE(Root Mean Squared Error)를 기준으로 예측력이 높은 상위 5개의 모형을 추출하였다. 추출된 모형에 검증용 데이터를 투입하여 예측값을 도출하고, 이를 실측값과 비교하여 모형의 성능을 검증하였다.

데이터 수집 및 처리

1. 데이터 수집

미세먼지 관련 빅데이터는 에어코리아 홈페이지에서 수집하였다. 진주시에 위치한 미세먼지 측정소는 표 2와 같이 총 4개소 있는데, 본 연구에서는 진주 상평산단을 대상으로 하기 때문에 측정소는 상평산단 내에 위치한 동진로 279(한국전력공사 진주지점)로 지정하였다. 수집한

TABLE 2. Fine dust measurement station location in Jinju

Name	Address	Measurement network
Daean-dong	1052, Jinju-daero, Jinju-si, Gyeongnam	City atmosphere
Sangbong-dong	12 Bibong-ro 85beon-gil, Jinju-si, Gyeongnam	City atmosphere
Sangdae-dong	270, Dongjin-ro, Jinju-si, Gyeongnam	City atmosphere
Jeongchon-myeon	1340 Yehari, Jeongchon-myeon, Jinju-si, Gyeongnam	City atmosphere

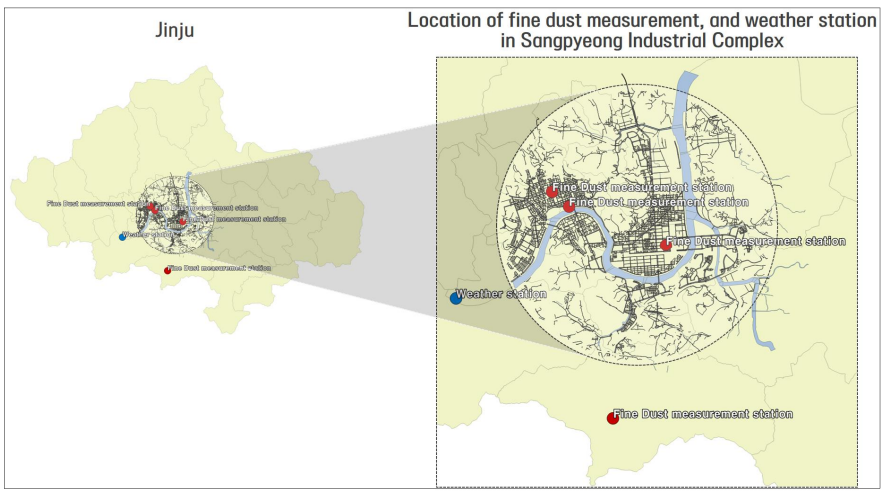

FIGURE 5. Location of fine dust measurement, and weather station in Sangpyeong industrial complex

자료는 미세먼지(PM10), 초미세먼지(PM2.5), 오존(O_3), 이산화질소(NO_2), 일산화탄소(CO), 아황산가스(SO_2)로 여섯 가지 측정값이다. 데이터 수집기간은 2019년 1월 1일 00시부터 2020년 12월 31일 24시까지로, 총 2년간의 자료이다.

기상 관련 빅데이터는 기상자료개방포털 홈페이지에서 수집하였다. 상평산단이 위치한 경상남도 진주시의 기상관측소는 단 1곳으로, 진주기상관측소의 위치는 경상남도 진주시 남강로 43이다. 수집한 기상자료는 기온, 강수량, 풍속, 풍향, 습도, 증기압, 이슬점 온도, 현지 기압, 해면 기압, 일조, 일사, 전운량, 시정 13가지 측정값이다. 데이터 수집기간은 2019년 1월 1일 00시부터 2020년 12월 31일 24시까지로 미세먼지 관련 데이터 수집기간과 동일하게 설정하였다.

진주시 상평산업단지의 위치와 미세먼지 측정소, 기상관측소의 위치는 그림 5와 같다. 본 연구에서 사용한 기상 관측소(경상남도 진주시 남강로 43)와 미세먼지 측정소(경남 진주시 동진로 279)의 직선거리는 약 7.7㎞이다.

2. 데이터 전처리

미세먼지와 기상 관련 빅데이터는 각각 정렬상태가 달라, 이를 가공하여 날짜 및 시간(년-월-일-시)을 기준으로 하나로 결합하였다(Lim *et al.*, 2018). 수집한 미세먼지 및 기상 데이터에는 결측값이 존재하는데, 강수량의 결측값은 0으로, 그 외의 기온, 풍속 등의 결측값은 평균값으로 대신하였다. 년-월-일-시에 따라 수집된 총 데이터의 수는 17,544개이다. 본 연구에서는 17,544개의 시간 자료를 731개의 일 자료로 변환하여 사용하였다. 24시간 데이터를 1일 데이터로 변환하기 위해서는 최소값, 최대값, 평균값 등 어떤 값을 사용할지 결정해야 한다.

TABLE 3. Data

Data	Variable	Missing value processing	Converted to daily data
Date data	year	average	year
	month	average	month
	day	average	day
Fine dust related data	PM10(㎍/㎥)	average	maximum
	PM2.5(㎍/㎥)	average	maximum
	O3(ppm)	average	maximum
	NO2(ppm),	average	maximum
	CO(ppm)	average	maximum
	SO2(ppm)	average	maximum
Weather related data	temperature fluctuations(℃)	average	temperature daily maximum −temperature daily maximum
	temperature(℃)	average	minimum
	precipitation(mm)	0	minimum
	wind speed(m/s)	average	maximum
	wind direction(16 directions)	average	average
	humidity(%)	average	maximum
	vapor pressure(hPa)	average	maximum
	dew point temperature(℃)	average	maximum
	local pressure(hPa)	average	maximum
	sea level pressure(hPa)	average	average
	duration of sunshine(hr)	average	maximum
	insolation(MJ/㎡)	average	maximum
	total cloud amount(10 quartile)	average	maximum
	visibility	average	minimum

이는 기온을 이용하여 일교차를, 풍향, 해면기압은 일중 평균치를, 강우량, 시정, 온도의 경우 일중 최소치를 사용하였고 나머지 예측인자들은 일 중 최대치를 사용하였다(Sohn and Kim, 2015). 데이터의 형태는 표 3과 같다. 이를 계절별로 나누어 훈련용 데이터와 검증용 데이터로 세 가지 비율 70:30, 80:20, 90:10으로 랜덤하게 2회 분류하였다.

머신러닝을 활용한 미세먼지 예측 모형 개발 및 예측력 평가

1. 다중회귀분석을 활용한 변수 선정

종속변수 PM10과 독립변수 22개로 1차 다중회귀분석을 실시한 결과는 표 4와 같다. 년, 월, 일, 오존, 일산화탄소, 기온, 강수량, 풍속, 현지 기압, 해면기압, 일조, 일사, 전운량 13개의 독립변수가 유의수준 5% 기준에서 유의하지 않은 것으로 판단되어 1차적으로 제거하였다.

제거 후 9개의 독립변수로 2차 다중회귀분석을 실시하여 다중공선성을 측정한 결과는 TABLE 5와 같다. 증기압, 이슬점 온도 2개의 독립변수가 공차가 0.1 미만, VIF(Variance Inflation Factor)가 10.0을 초과하여 다시 제거하였다.

최종적으로 남은 7개의 독립변수 PM2.5, NO_2, SO_2, 일교차, 풍속, 습도, 시정과 종속변수 PM10의 2년간의 추이는 그림 6과 같다. 8개의 변수 대부분은 계절에 따라 차이를 보이며, 어느 정도 일정한 추세를 가지는 것으로 나타났다. 미세먼지(PM10)와 초미세먼지(PM2.5)는 수치폭의 차이는 있지만 유사한 형태로 변동하고 있으며, NO_2는 겨울에 높아지며, SO_2는 계절과 상관없이 증감을 반복하는 것으로 분석되었다. 일교차는 여름과 가을에 작고, 풍속은

TABLE 4. First multiple regression analysis result

Independent variable	B value	t value	p value
year	−1.658	−1.116	.265
month	−0.510	−1.915	.056
day	−0.029	−0.343	.732
PM2.5_MAX	1.107	13.364	.000
O3_MAX	13.465	0.532	.595
NO2_MAX	357.371	4.265	.000
CO_MAX	5.315	1.240	.215
SO2_MAX	373.513	2.312	.021
temperature fluctuations	1.148	2.627	.009
temperature_MIN	1.011	1.958	.051
precipitation_MIN	−58.978	−0.300	.764
wind speed_MAX	3.189	3.444	.001
wind direction_AVERAGE	0.000	0.010	.992
humidity_MAX	−0.441	−2.760	.006
vapor pressure_MAX	−3.207	−7.861	.000
dew point temperature_MAX	2.323	4.099	.000
local pressure_MAX	−0.158	−0.240	.810
sea level pressure_AVERAGE	0.239	0.378	.705
duration of sunshine_MAX	1.171	0.253	.800
insolation_MAX	−2.795	−1.302	.193
total cloud amount_MAX	−0.409	−1.257	.209
visibility_MIN	−0.005	−2.460	.014

TABLE 5. Second multiple regression analysis result

Independent variable	Tolerance	VIF
PM2.5_MAX	.461	2.171
NO_2_MAX	.456	2.193
SO_2_MAX	.820	1.219
temperature fluctuations	.442	2.264
wind speed_MAX	.851	1.175
humidity_MAX	.287	3.488
vapor pressure_MAX	.046	21.928
dew point temperature_MAX	.036	27.515
visibility_MIN	.387	2.582

봄과 가을이 높은 편이다. 습도는 겨울에 낮아지며, 시정은 계절과 상관없이 증감을 반복하는 특징이 있다.

7개의 독립변수 PM2.5, NO_2, SO_2, 일교차, 풍속, 습도, 시정을 대상으로 종속변수 PM10에 대한 3차 다중회귀분석을 실시하였다. 분석결과는 표 6, 7과 같다.

3차 다중회귀분석 후 R제곱 값은 0.541으로,

TABLE 6. Analysis of variance

model	Sum of squares	Degrees of freedom	Mean squared	F value	p value
Regression	350124.084	7	50017.726	121.886	0.000
Residual	296693.864	723	410.365		
Total	646817.948	730			

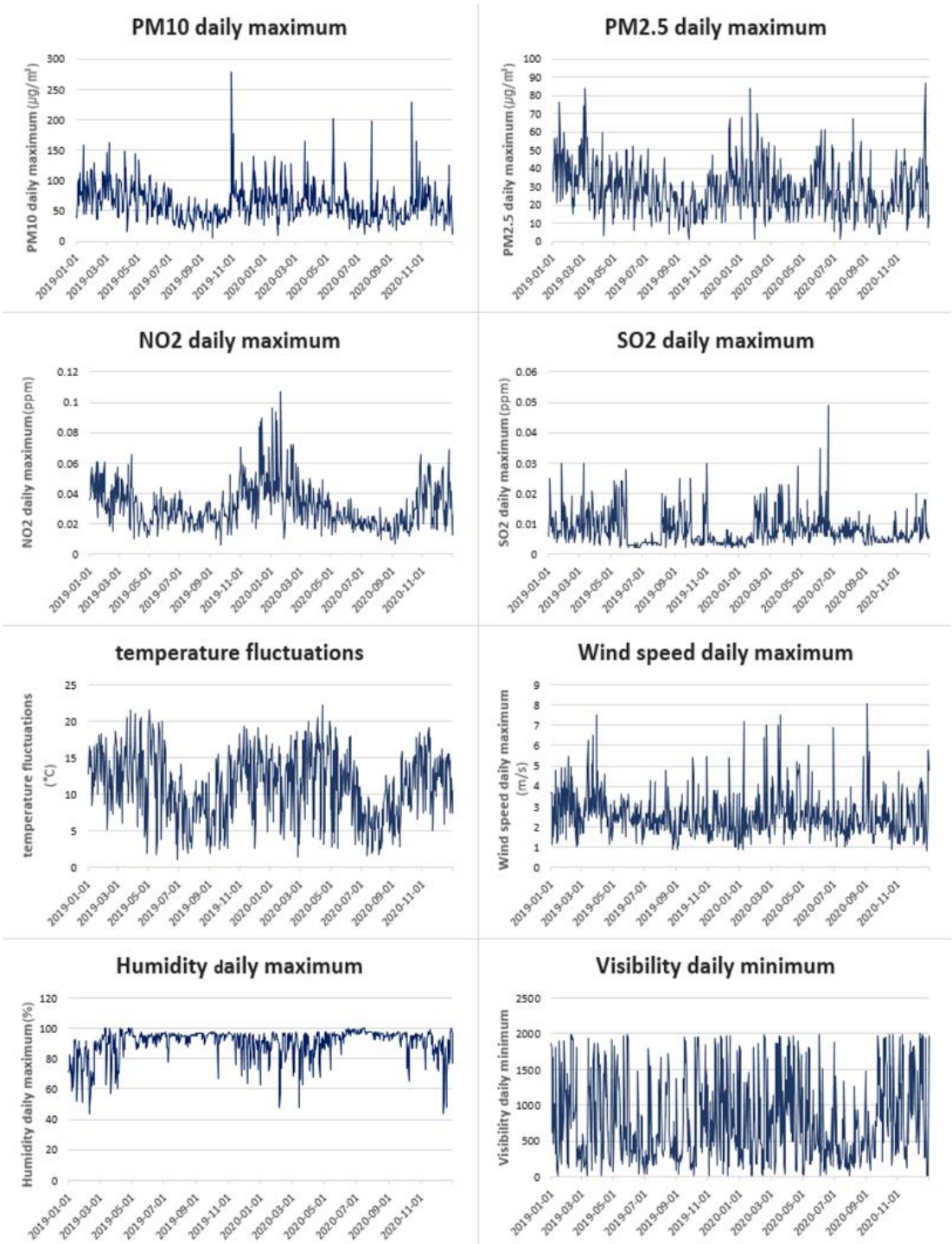

FIGURE 6. Graph of dependent and independent variables

TABLE 7. Third multiple regression analysis result

Model	R	R^2	Adjusted R^2	Standard error of the estimate
Regression	0.736	0.541	0.537	20.257

$$PM10\_MAX=1.068(PM2.5\_MAX)+334.002(NO_2\_MAX)+441.206(SO_2\_MAX)+1.158(temperature\ fluctuations)+3.133(wind\ speed\_MAX)-0.056(humidity\_MAX)-0.005(visibility\_MIN) \quad (1)$$

최종적으로 도출된 회귀식의 설명력은 약 54.1%이다.

F값은 1717.398이며, 유의확률은 0.000으로 유의수준 5%를 충족하여 최종 도출된 회귀식이 유의함을 알 수 있다. 최종적으로 도출된 회귀식은 식 1과 같다.

2. 데이터세트

연구 대상지의 PM10 농도는 계절별로 차이를 보인다. 그림 7과 같이, 계절별 평균 PM10 농도는 봄과 겨울, 여름과 가을이 각각 유사하며, 그 수치는 봄, 겨울이 63~77ppm, 여름, 가을이 49~58ppm 정도로 봄과 겨울에 미세먼지 농도가 더 높음을 알 수 있다. 따라서 다중회귀분석 후 도출된 최종 데이터세트를 계절별로 분류하였다. 계절별 데이터세트는 봄 184개, 여름 184개, 가을 182개, 겨울 181개로, 이를 다시 70:30, 80:20, 90:10 세 가지 비율로 랜덤하게 2회씩 분류하여 총 24개의 훈련용 데이터와 24개의 검증용 데이터를 구축하였다.

3. 예측모형 개발 및 예측력 평가

1) RMSE를 통한 예측력 평가

훈련용 데이터 각각을 19개의 머신러닝 회귀학습기 모형으로 학습하여 각각의 경우에서 예측력이 높은 상위 5개의 모형을 추출하였다. 모형의 예측력은 평균제곱근 오차인 RMSE(Root Mean Squared Error)를 통해서 판단하였으며, RMSE 식은 아래와 같다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\tilde{y_i})^2} \quad (2)$$

y_i: 실제 미세먼지 농도

$\tilde{y_i}$: 예측한 미세먼지 농도

n: 데이터 건수

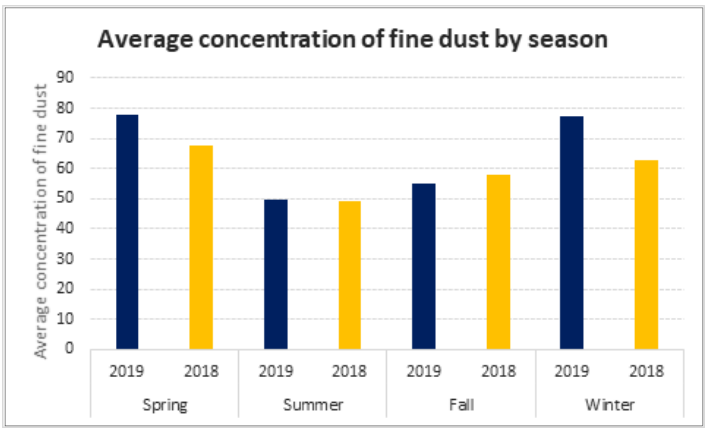

FIGURE 7. Average concentration of PM10 by season

RMSE 값이 0에 가까울수록 오차의 정도가 적은, 즉 예측력이 높은 모형이라고 할 수 있다. 머신러닝 훈련 결과 RMSE값은 FIGURE 8과 같다. 예측력이 높은 모형은 중복적으로 유사하게 나타났으며, 크게 세 가지이다. 첫째는 선형회귀모형, 로버스트 선형회귀모형, 단계적 선형회귀모형과 같은 선형회귀모형이다. 선형회귀모형은 선형회귀분석을 통해 종속변수와 독립변수 간의 관계를 선형식으로 나타내는 모형이다. 두 번째는 유리 2차 GPR, 제곱지수 GPR, 매턴 5/2 GPR과 같은 가우스 과정 회귀모형이다. 가우스 과정 회귀모형은 일반적인 머신러닝 기법과 다르게 예측이 확률기반이기 때문에 학습에 사용되는 데이터를 이용해 예측값이 존재할 영역을 계산하고 신뢰구간을 알 수 있다(An and Ryoo, 2016). 즉, 데이터를 기반으로 추정된 예측 값의 분포를 통해 불확실성이 고려된 예측을 이끌어낼 수 있다(Jeong and Park, 2017). 세 번째는 선형 SVM, 성긴 가우스 SVM과 같은 서포트 벡터 머신이다. 서포트 벡터머신은 훈련 데이터들을 학습시켜 최적의 초평면을 찾는다. 즉, 서로 다른 범주에 속한 관측치 사이에 간격이 최대가 되는 선을 찾는다(Sung et al., 2020). 즉, 집단으로 분류할 수 있는 기준을 바탕으로 새로운 데이터가 주어졌을 때, 어느 집단으로 분류하는지를 판단하는 알고리즘이라고 할 수 있다(Jung *et al.*, 2020).

훈련용 데이터와 시험용 데이터의 비율별로 가장 예측력이 우수한 경우는 다음과 같다. 봄, 여름, 겨울은 70:30, 가을은 80:20의 비율로 훈련용 데이터와 시험용 데이터를 나눌 때, 예측력이 가장 높은 것으로 나타났다. 이와 같이 대부분의 경우에서 훈련용 데이터와 시험용 데이터의 비율을 90:10으로 나누는 것 보다, 70:30, 80:20으로 나누는 경우 모형의 예측력이 더욱 향상됨을 알 수 있다. 즉, 훈련용 데이터의 양과 머신러닝 모형의 예측력이 비례하지 않음을 알 수 있다.

최종적으로 계절별로 가장 예측력이 우수한 모형을 살펴보면, 봄과 여름은 70% 비율의 학습용 데이터로 학습된 로버스트 선형회귀모형이, 가을은 80% 비율의 학습용 데이터로 학습된 2차 SVM모형이, 겨울은 70% 비율의 학습용 데이터로 학습된 선형 SVM모형이 가장 예측력이 우수한 것으로 나타났다.

2) 실측치와 예측치의 비교를 통한 예측력 평가

봄은 70:30의 비율로 학습한 로버스트 선형회귀모형이 RMSE 기준으로 예측력이 가장 우수한 것으로 나타났다. 해당 모형에 검증용 데이터를 넣어 예측치를 도출한 결과, |실측치−예측치|의 최소값은 0.14이었으며, 이 경우는 실측치가 71ppm인 실측치가 낮은 날이었다. |실측치−예측치|의 최대값은 145.97이며, 예측치가 실측치와 가장 큰 차이를 보인 경우는 실측치가 201ppm로 가장 높게 나타난 날이었다. |실측치−예측치|의 평균값은 17.54이며, 봄 데이터셋의 30% 비율 검증용 데이터 총 55개 중 41개가 평균값 이하의 차이를 보였다. 즉, 실측치와 예측치의 차이가 큰 경우는 많지 않았다.

여름은 70:30의 비율로 학습한 로버스트 선형 회귀모형이 RMSE 기준으로 예측력이 가장 우수하였다. 해당 모형에 검증용 데이터를 넣어 예측치를 도출한 결과, |실측치−예측치|의 최소값은 0.05이었으며, 이 경우는 실측치가 27ppm로 낮은 날이었다. |실측치−예측치|의 최대값은 166.29로서, 예측치가 실측치와 가장 큰 차이를 보인 경우는 실측치가 197ppm로 가장 높은 날이었다. |실측치−예측치|의 평균값은 9.14로서, 여름 데이터셋의 30% 비율 검증용 데이터 총 55개 중 40개가 평균값 이하의 차이를 보였다. 즉, 실측치와 예측치의 차이가 큰 경우는 많지 않았다.

가을은 80:20의 비율로 학습한 2차 SVM모형이 RMSE 기준으로 예측력이 가장 우수하였다. 해당 모형에 검증용 데이터를 넣어 예측치를 도출한 결과, |실측치−예측치|의 최소값은 0.08이었으며, 이 경우는 실측치가 69ppm로 비교적 높으며, 봄, 여름, 겨울의 최적 모형과 차이를 보였다. |실측치−예측치|의 최대값은 182.48로서 예측치가 실측치와 가장 큰 차이를

보인 경우는 실측치가 229ppm로 가장 높게 나타난 날이었다. |실측치-예측치|의 평균값은 16.01이며, 가을 데이터셋의 20% 비율 검증용 데이터 총 36개 중 30개의 경우가 평균값 이하의 차이를 보였다. 즉, 실측치와 예측치의 차이가 큰 경우는 많지 않았다.

겨울은 70:30의 비율로 학습한 선형 SVM모형이 RMSE 기준으로 예측력이 가장 우수하였다. 해당 모형에 검증용 데이터를 넣어 예측치를 도출한 결과, |실측치-예측치|의 최소값은 0.22로 나타났으며, 이 경우는 실측치가 39ppm로 가장 낮게 나타난 날이었다. |실측치-예측

TABLE 8. Machine learning results

Train : Test		Spring	Summer	Fall	Winter
Train:Test =70:30	1	Robust Regression (19.604)	Linear Regression (22.509)	Quadratic SVM (21.824)	Linear SVM (13.969)
		Linear Regression (19.663)	Robust Regression (22.512)	Linear SVM (22.543)	Squared Exponential GPR (13.978)
		Linear SVM (19.73)	Squared Exponential GPR (22.678)	Robust Regression (22.553)	Rational Quadratic GPR (13.978)
		Squared Exponential GPR (19.979)	Rational Quadratic GPR (22.678)	Stepwise Regression (23.017)	Robust Regression (13.985)
		Rational Quadratic GPR (19.979)	Matern 5/2 GPR (22.701)	Linear Regression (23.363)	Matern 5/2 GPR (14.008)
	2	Robust Regression (17.416)	Robust Regression (11.234)	Cubic SVM (27.391)	Matern 5/2 GPR (16.018)
		Linear Regression (17.441)	Linear SVM (11.395)	Robust Regression (27.493)	Squared Exponential GPR (16.021)
		Linear SVM (17.595)	Linear Regression (11.594)	Quadratic SVM (27.517)	Rational Quadratic GPR (16.021)
		Squared Exponential GPR (18.011)	Stepwise Regression (11.631)	Linear SVM (27.603)	Coarse Gaussian SVM (16.047)
		Rational Quadratic GPR (18.011)	Coarse Gaussian SVM (12.239)	Bagged Tree (28.337)	Linear SVM (16.06)
Train:Test =80:20	1	Quadratic SVM (23.744)	Linear SVM (16.889)	Quadratic SVM (27.166)	Quadratic SVM (15.431)
		Squared Exponential GPR (24.264)	Robust Regression (16.976)	Linear SVM (27.786)	Linear Regression (15.524)
		Rational Quadratic GPR (24.264)	Stepwise Regression (17.112)	Robust Regression (27.914)	Robust Regression (15.557)
		Matern 5/2 GPR (24.321)	Linear Regression (17.272)	Bagged Tree (28.625)	Exponential GPR (15.803)
		Linear SVM (24.379)	Coarse Gaussian SVM (17.473)	Coarse Gaussian SVM (28.805)	Squared Exponential GPR(15.962)
	2	Boosted Tree (22.177)	Linear SVM (16.981)	Quadratic SVM (19.99)	Quadratic SVM (16.094)
		Linear Regression (22.229)	Robust Regression (17.045)	Linear SVM (21.403)	Robust Regression (16.172)
		Robust Regression (22.278)	Stepwise Regression (17.079)	Robust Regression (21.44)	Linear Regression (16.43)
		Stepwise Regression (22.397)	Linear Regression (17.146)	Stepwise Regression (21.451)	Stepwise Regression (16.434)
		Matern 5/2 GPR (22.48)	Matern 5/2 GPR (17.24)	Cubic SVM (21.696)	Squared Exponential GPR (16.608)

TABLE 8. Continued

Train : Test		Spring	Summer	Fall	Winter
Train:Test =90:10	1	Robust Regression (22.532)	Robust Regression (16.23)	Quadratic SVM (25.627)	Exponential GPR (15.246)
		Linear SVM (22.756)	Linear SVM (16.261)	Cubic SVM (25.791)	Rational Quadratic GPR (15.667)
		Linear Regression (22.86)	Stepwise Regression (16.517)	Bagged Tree (26.534)	Squared Exponential GPR (15.687)
		Matern 5/2 GPR (22.862)	Quadratic SVM (16.649)	Stepwise Regression (26.746)	Linear SVM (15.732)
		Bagged Tree (22.901)	Coarse Gaussian SVM (16.709)	Robust Regression (26.793)	Matern 5/2 GPR (15.733)
	2	Linear Regression (22.509)	Robust Regression (16.235)	Quadratic SVM (25.522)	Quadratic SVM (15.395)
		Robust Regression (22.512)	Stepwise Regression (16.241)	Bagged Tree (26.205)	Robust Regression (15.6)
		Squared Exponential GPR (22.678)	Linear SVM (16.25)	Stepwise Regression (26.366)	Squared Exponential GPR (15.708)
		Rational Quadratic GPR (22.678)	Linear Regression (16.603)	Robust Regression (26.447)	Rational Quadratic GPR (15.708)
		Matern 5/2 GPR (22.701)	Coarse Gaussian SVM (16.883)	Linear SVM (26.467)	Matern 5/2 GPR (15.772)

치| 의 최대값은 68.31으로서, 예측치가 실측치와 가장 큰 차이를 보인 경우는 실측치가 129ppm로 가장 높은 날이었다. |실측치−예측치| 의 평균값은 13.16으로서, 겨울 데이터셋의 30% 비율 검증용 데이터 총 55개 중 33개가 평균값 이하의 차이를 보였다. 즉, 겨울의 최적 모형은 사계절 중 |실측치−예측치| 최대값이 가장 낮게 나타났지만, |실측치−예측치| 평균값 이상의 차이를 보이는 경우가 40%일 정도로 사계절 중 가장 많은 것으로 분석되었다.

각 계절별로 구축한 미세먼지 예측 모형의 예측치와 실측치의 차이는 평균적으로 크지 않았다. 특히 실측치가 비교적 낮은 경우에는 실측치와 예측치의 차이가 적게 나타나 예측력이 높았다. 하지만 사계절 모두, 실측치가 큰 경우에서 예측치의 차이가 많고, 예측력도 다소 낮았다. 따라서 예측 모형을 계절과 별개로 실측치를 적당한 수치 기준으로 분류하여 각각 모형을 구축해 보는 것도 모형의 예측력을 높이는 방법이 될 수 있을 것이다. 또한 본 연구에서 구축한 독립변수(기상 관련 데이터 4개, 일교차, 풍속, 습도, 시정, 미세먼지 관련 데이터 3개, PM2.5, NO_2, SO_2)는 미세먼지의 농도가 낮은 경우에는 미세먼지 농도에 영향력이 높지만, 반대의 경우에는 비교적 영향력이 낮으며, 이 경우 다른 유의한 변수를 탐색하여 모형의 예측력을 높일 필요가 있을 것이다.

결 론

연구사례가 많지 않은 지방 산단인 진주상평산업단지를 대상으로 기상과 미세먼지 관련 빅데이터로 미세먼지(PM10) 농도를 예측해보았다. 2019년 1월 1일부터 2020년 12월 31일까지 총 17,544시간 데이터를 수집한 후, 731개의 일 데이터로 변환하여 데이터셋을 구축하였다. 구축된 데이터셋은 종속변수 미세먼지(PM10) 1개와 22개의 독립변수로 구성되었다.

이들 자료는 다중회귀분석을 통해 종속변수에 유의한 영향을 주는 독립변수를 탐색하였다. 다중회귀분석은 총 3회 실시하였으며, 1차 다중회귀분석 결과 년, 월, 일, 오존, 일산화탄소, 기온, 강수량, 풍속, 현지 기압, 해면기압, 일조, 일사, 전운량 13개의 독립변수가 유의하지 않아 1차

적으로 제거하였다. 다음으로 2차 다중회귀분석에서 다중공선성을 기준으로 증기압, 이슬점 온도 2개의 독립변수를 제거하였다. 마지막 3차 다중회귀분석에서 최종 회귀식을 도출하였다. 결과적으로 최종 데이터셋은 미세먼지(PM10) 1개와 독립변수 PM2.5, NO_2, SO_2, 일교차, 풍속, 습도, 시정 7개로 구성되었다.

최종 데이터셋은 미세먼지 농도가 계절별로 차이가 있기 때문에 계절별로 분리하였다. 계절별 데이터셋은 다시 70:30, 80:20, 90:10 세 가지 비율로 랜덤하게 2회씩 분류하여 훈련용 데이터와 검증용 데이터를 구축하였다. 훈련용 데이터는 각각 19개의 머신러닝 회귀학습기 모형으로 학습하여 예측력이 높은 모형을 추출하였다. 예측력은 RMSE로 판단하였으며, 예측력이 높은 모형은 선형회귀모형, 로버스트 선형회귀모형, 단계적 선형회귀모형과 같은 선형회귀모형과 유리 2차 GPR, 제곱지수 GPR, 매턴 5/2 GPR과 같은 가우스 과정 회귀모형, 선형 SVM, 성긴 가우스 SVM과 같은 서포트 벡터머신으로 나타났다.

또한, 훈련용 데이터의 비율 측면에서는 훈련용 데이터와 시험용 데이터의 비율을 90:10으로 나누는 것 보다, 70:30, 80:20으로 나누는 경우, 모형의 예측력이 더욱 향상됨을 알 수 있었다. 이에 훈련용 데이터의 비율이 높아지면 더 많은 데이터로 학습되어 예측력도 우수해질 것이라 예상했으나, 훈련용 데이터의 비율과 예측력은 비례하지 않음을 알 수 있었다. 예측력이 우수한 모형을 검증용 데이터로 검증해본 결과, 예측 모형의 예측치와 실측치 차이는 크지 않은 것으로 나타났다. 특히 비교적 실측치가 낮은 경우에는 실측치와 예측치의 차이가 적게 나타나 예측력이 높았지만 실측치가 높은 경우에서는 실측치와 예측치의 차이가 크게 나타나 예측력이 낮아짐을 알 수 있었다.

이상과 같이 미세먼지 예측모형을 개발해 보았지만, 향후 스마트도시 통합관제센터나 일부 지자체에서 설치 계획 중인 데이터 허브에 기상데이터와 관련 도시 빅데이터를 결합함으로써 보다 체계적이고 정밀한 미세먼지 예측 서비스로 개발이 가능할 것이다. 그 결과는 시민의 건강 증진과 노후산단이 스마트그린산단으로 개조될 수 있는 시작점이 될 수 있을 것이며, 미래형 스마트도시를 앞당기는데 기여할 수 있을 것이다. KAGIS

REFERENCES

AirKorea. 2021. Data Retrieve. https://www.airkorea.or.kr. (Accessed January 25, 2021).

An, M.H. and Ryoo, M.H. 2016. Modeling Stochastic Volatility Using Gaussian Processes. The Korean Journal Of Financial Engineering 2(0):101-113 (안명호, 유미현. 2016. Gaussian Process를 이용한 Stochastic Volatility Modeling. 한국금융공학회 2(0):101-113).

Bae, S.W. and Yu, J.S. 2018. Predicting the Real Estate Price Index Using Machine Learning Methods and Time Series Analysis Model. Korean Association For Housing Policy Studies 26(1):107-133 (배성완, 유정석. 2018. 머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측. 한국주택학회 26(1):107-133).

Cha, J.W. and Kim, J.Y. 2018. Development of Data Mining Algorithm for Implementation of Fine Dust Numerical Prediction Model. Korea Institute of information and Communication Engineering 22(4):595-601 (차진욱, 김장영. 2018. 미세먼지 수치예측 모델 구현을 위한 데이터마이닝 알고리즘 개발. 한국정보통신학회 22(4):595-601).

Cho, K.W., Jeong, Y.J., Lee, J.S. and Oh, C.H. 2019. PM10 Particulate Matters Concentration Prediction using Korea Institute of information and Communication Engineering 23(2):632-634 (조경우, 정용

진, 이종성, 오창헌. 2019. LSTM을 이용한 PM10 미세먼지 농도 예측. 한국정보통신학회 23(2):632-634).

Choi, C.H., Kim, J.S., Kim, J.H., Kim, H.Y., Lee, W.J. and Kim, H.S. 2017. Development of Heavy Rain Damage Prediction Function Using Statistical Methodology. Korean Society of Hazard Mitigation 17(3):331-338 (최창현, 김종성, 김정환, 김한용, 이우주, 김형수. 2017. 통계적 방법론을 이용한 호우피해예측함수 개발. 한국방재학회 17(3):331-338).

Heo, S.Y., Kim, J.Y. and Moon, T.H. Predicting Crime Risky Area Using Machine Learning. Journal of the Korean Association of Geographic Information Studies 2018. 21(4):64-80 (허선영, 김주영, 문태헌. 2018. 머신러닝기반 범죄발생 위험지역 예측. 한국지리정보학회 21(4):64-80).

Jeon, S.H. and Son, Y.S. 2018. Prediction of fine dust PM10 using a deep neural network model. The Korean Journal of Applied Statistics 31(2):265-285 (전성현, 손영숙. 2018. 심층 신경망모형을 사용한 미세먼지 PM10의 예측. 응용통계연구 31(2):265-285).

Jeong, Y.H. and Park, J.K. 2017. Energy Storage system Strategy under Gaussian Process Regression. The Korean Institute of Industrial Engineers 2017(04):2690-2695 (정요한, 박진규. 2017. 예측모델 (Gaussian Process Regression)을 통한 에너지저장시스템 운영전략. 대한산업공학회 춘계학술대회 논문집 2017(04):2690-2695).

Jung, Y.J., Cho, K.W., Lee, J.S. and Oh, C.H. 2020. PM10 Binary Classification Model based on SVM Algorithm. Korea Institute of information and Communication Engineering 24(1):308-310 (정용진, 조경우, 이종성, 오창헌. 2020. SVM 알고리즘 기반의 PM10 이진 분류 모델. 한국정보통신학회 24(1):308-310).

Kim, B.H., Lee, H.Y. and Lee, S.M. 2019. The Inutitute of Electronic and information Engineers 6(11):53-60 (김병희, 이희용, 이순미. 2019. 무인항공기 대기측정센서 데이터와 LSTM 모델을 활용한 미세먼지 (PM10) 농도 예측. 대한전자공학회 46(11):53-60).

Kim, H. 2020. The Prediction of PM2.5 in Seoul through XGBoost ensemble. Journal of the Korean Data Analysis Society 22(4):1661-1671 (김혁. 2020. XGBoost 앙상블에 의한 서울시 초미세먼지 예측. 한국자료분석학회 22(4):1661-1671).

Kim, J.S., Choi, C.H., Kim, D.H., Lee, M.J. and Kim, H.S. 2017. Development of Heavy Rain Damage Prediction Function Using Artificial Neural Network and Multiple Regression Model. Korean Society of Hazard Mitigation 17(6):73-80 (김종성, 최창현, 김동현, 이명진, 김형수. 2017. 인공신경망과 다중회귀모형을 이용한 호우피해 예측함수 개발. 한국방재학회 17(6):73-80).

Kyung, S.Y., Kim, Y.S., Kim, W.J., Park, M.S., Song. J.W., Yum, H.K., Yoon, H.G., Rhee, C.K. and Jeong, S.H. 2015. Guideline for the prevention and managementof particulate matter/Asian dust particleinduced adverse health effect on the patients withpulmonary diseases. Journal of the Korean Medical Association 58(11):1060-1069 (경선영, 김영삼, 김우진, 박무석, 송진우, 염호기, 윤형규, 이진국, 정성환. 2015. 미세먼지/황사 건강피해 예방 및 권고지침: 호흡기질환. 대한의사협회지 58(11):1060-1069).

Lee, A.R. and Jeong, S.J. 2019. Korean Meteorological Society 2019(10):357-357 (이아름, 정수종. 2019. 인공위성 데이터를 활용한 딥러닝 기반의 서울 내 미세먼지 농도 예측. 한국기상학회 학술대회 논문집 2019(10):357-357).

Lee, D.W. and Lee, S.W. 2020. Hourly Prediction of Particulate Matter (PM2.5) Concentration Using Time Series Data and Random Forest. Korea Information Processing Society 9(4):129-136 (이득우, 이수원. 2020. 시계열 데이터와 랜덤 포레스트를 활용한 시간당 초미세먼지 농도 예측. 한국정보처리학회 9(4):129-136).

Lee, Y.S. and Moon, P.J. 2017. A Comparison and Analysis of Deep Learning Framework. Korea Institute of Electronic Communication Sciences 12(1):115-122 (이요섭, 문필주. 2017. 딥 러닝 프레임워크의 비교 및 분석. 한국전자통신학회 12(1):115-122).

Lim, J.M. 2019. An Estimation Model of Fine Dust Concentration Using Meteorological Environment Data and Machine Learning. Journal of Information Technology Services 18(1):173-186 (임준묵. 2019. 기상환경데이터와 머신러닝을 활용한 미세먼지농도 예측 모델. 한국IT서비스학회지 18(1):173-186).

Lim, J.M., Ko, S.H. and Kim, J.W. 2018. An estimation model of fine dust concentration using weather data and machine learning. Korea Society of IT Services 2018:691-694 (임준묵, 고선호, 김제완. 기상데이터와 머신러닝을 활용한 미세먼지농도 예측 모델. 한국IT서비스학회 2018:691-694).

Ministry of Environment. 2017. Comprehensive measures for fine dust management. pp.1-37 (환경부. 2017. 미세먼지 관리 종합대책. pp.1-37).

Ministry of Trade, Industry and Energy. 2020. Implementation strategies of Smart Green industrial complex. pp.1-27 (산업통상자원부. 2020. 스마트그린산단 실행 전략. pp.1-27).

Oh, J.M., Shin, H.S., Shin, Y.S. and Jeong. H.C. 2017. Forecasting the Particulate Matter in Seoul using a Univariate Time Series Approach. Korea Information Processing Society 19(5):2457-2468 (오종민, 신현수, 신예슬, 정형철. 2017. 시계열 분석을 활용한 서울시 미세먼지예측. 한국자료분석학회 19(5):2457-2468).

Open MET Data Portal. 2021. Data. https://data.kma.go.kr. (Accessed January 25, 2021).

Sohn, K.T. and Kim, D.H. 2015. Development of statistical forecast model for PM10 concentration over Seoul. Journal of the Korean data & information science society 26(2):289-299 (손건태, 김다홍. 2015. 서울지역 PM10 농도 예측모형 개발. 한국데이터정보과학회 26(2):289-299).

Sung, J.H. and Cho, Y.S. 2019. Machine Learning Approach for Pattern Analysis of Energy Consumption in Factory. Korea Information Processing Society 8(4):87-92 (성종훈, 조영식. 2019. 머신러닝 기법을 활용한 공장 에너지 사용량 데이터 분석. 한국정보처리학회 8(4):87-92).

Sung, S.H., Kim, S.J. and Ryu, M.H. 2020. A Comparative Study on the Performance of Machine Learning Models for the Prediction of Fine Dust: Focusing on Domestic and Overseas Factors. Korea Society Of Innovation 15(4):339-357 (성상하, 김상진, 류민호. 2020. 미세먼지 예

측을 위한 기계학습 모델 간 성능 비교 연구: 국내 발생 데이터를 중심으로. 한국혁신학회 15(4):339-357).

Yeo, M.S. and Kim, J.H. 2019. The Society Of Air-Conditioning And Refrigerating Engineers Of Korea 48(12):44-50 (여명석, 김지혜. 2019. 어린이집 보육실의 실내 활동 분석을 통한 실내 미세먼지 발생률 예측. 대한설비공학회 48(12):44-50).

Yoon, C.J., Ko, H.G., Bang, M.S. and Kwon, H.B. 2018. An Analysis of the Rail Wear Measurements for the Prediction of Particulate Matter Emission in Urban Railway. Journal of Korean Society for Urban Railway 6(4):339-350 (윤천주, 고희규, 방명석, 권혁빈. 2018. 도시철도 미세먼지 발생량 예측을 위한 레일 마모량 분석. 한국도시철도학회 6(4):339-350).