정책지원과제(기후변화) 2022-06

인공지능기법을 활용한 측정소 미설치 지역 PM_{10} 농도 예측

기후변화대응연구센터

충청남도 서해안기후환경연구소

2022. 12

충남연구원
ChungNam institute

인공지능기법을 활용한 측정소 미설치 지역 PM10 농도 예측

2022. 12

충남연구원 ChungNam Institute
충청남도 서해안기후환경연구소 Seohaean Research Institute

Contents

목차 ····· i
그림목차 ····· ii

1장 연구개요 ····· 2
1. 연구배경 및 목적 ····· 2
2. 연구방법 ····· 4

2장 인공지능 기법을 활용한 대기질 예측 사례 문헌조사 ····· 6
1. 대기오염물질 예보시스템 ····· 6
2. 지하역사 미세먼지 센서 오류 감지 및 데이터 복원 ····· 7
2. 대기오염물질 자료 관리 ····· 8

3장 화력발전소 인간 민간측정망 데이터 전처리 및 통계분석 ····· 10
1. 마을대기측정망 개요 ····· 10
2. 데이터 전처리 ····· 12
3. 통계분석 ····· 13

4장 측정소 미설치 지역의 미세먼지(PM_{10}) 농도 예측 ····· 23
1. 적용한 인공지능 기법 상세 ····· 23
2. 대기오염물질 예측 결과 ····· 29
3. 활용방안 제안 ····· 33

참고문헌 ····· 34

그림목차

[그림 1] 연료별 국내 에너지 생산량(1961년~2019년) 및 기여도 ·····2
[그림 2] 국내 지역별 에너지 생산 추이 및 석탄화력발전 현황 ·····3
[그림 3] PM10 농도 예측 프로세스 ·····4
[그림 4] 부산시보건환경연구원에서 사용중인 PM2.5 예측모델 ·····6
[그림 5] IoT센서 이상데이터 판별 프로세스 및 자동제어시스템 운영방식 ··7
[그림 8] 인공지능을 활용한 데이터 학습 및 예측 사례 ·····7
[그림 6] A사의 이상감지모델 프로세스 ·····8
[그림 7] 충청남도 화력발전소와 대형배출시설 위치 ·····10
[그림 8] 2021년, 2022년 마을대기측정망 평균 유효가동률 변화 ·····11
[그림 9] 결측치 분포도 ·····12
[그림 10] 국가대기측정망(당진시청사)와 마을대기측정망(중흥) 자료 산점도 ···13
[그림 11] 당진지역 측정소별 연평균 아황산가스(SO2) 농도> ·····14
[그림 12] 당진지역 측정소별 월별 SO2 농도 변화> ·····15
[그림 13] 당진지역 측정소별 연평균 이산화질소(NO2) 농도> ·····15
[그림 14] 당진지역 측정소별 월별 NO2 농도 변화> ·····16
[그림 15] 당진지역 측정소별 연평균 오존(O3) 농도> ·····17
[그림 16] 당진지역 측정소별 월별 O3 농도 변화> ·····17
[그림 17] 당진지역 측정소별 연평균 일산화탄소(CO) 농도> ·····18
[그림 18] 당진지역 측정소별 월별 CO 농도 변화> ·····18
[그림 19] 당진지역 측정소별 연평균 미세먼지(PM10) 농도> ·····19
[그림 20] 당진지역 측정소별 월별 PM10 농도 변화> ·····20
[그림 21] 당진지역 측정소별 연평균 초미세먼지(PM2.5) 농도> ·····20
[그림 22] 당진지역 측정소별 월별 PM2.5 농도 변화> ·····21
[그림 23] RNN의 기본원리 ·····23
[그림 24] Schematic diagram of backpropagation by chain-rule ·····25
[그림 25] LSTM basic structure ·····26
[그림 26] Unit cell structure of Long-Short Term Memory ·····27
[그림 27] 당진지역 측정소 대기오염물질 농도 예측 결과 ·····30

[그림 28] 미세먼지 고농도 이벤트 발생시 실제값과 예측값의 비교 ·····31
[그림 29] 24시간 후 대기오염물질 농도 예측 결과 ·····31
[그림 15] 생활환경을 고려한 교실 내 공기질 관리방안 제시 ·····32

제 1 장

연구개요

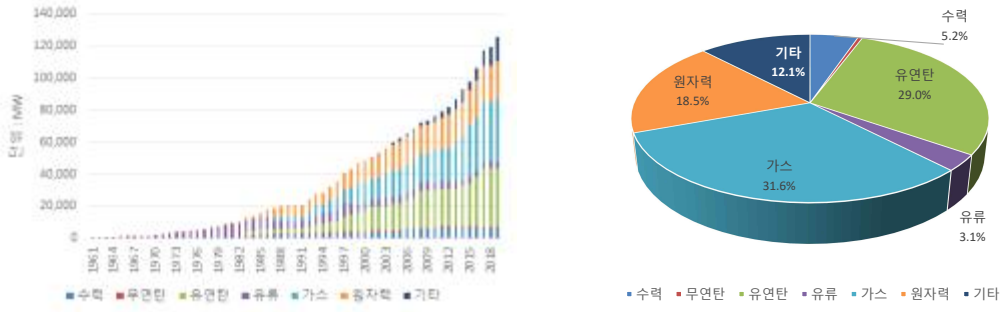
1. 연구배경 및 목적
2. 연구방법

제 1 장

연구개요

1. 연구배경 및 목적

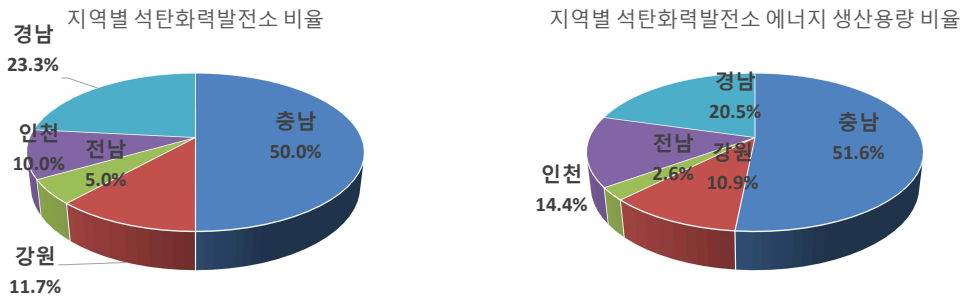
국내에서는 생산량 증가와 산업계의 꾸준한 발전으로 인하여 전력소비량이 매년 우상향하고 있다. 국내 에너지통계가 집계되기 시작한 1961년 이후 석탄의 에너지생산 기여율은 꾸준히 유지되었고, 에너지 생산량은 1961년 223 MW에서 2019년 37,003 MW까지 166배가 증가하였다. 2019년 에너지원별 생산량에서도 가스의 31.6%에 이어 유연탄이 29.0%로 2위를 차지할 정도로 아직까지 높은 기여도를 보이고 있다.



[그림 1] 연료별 국내 에너지 생산량 및 기여도

충청남도에는 2022년 기준으로 석탄을 사용하여 발전하는 석탄화력발전소가 전국의 약 50%가 위치하고 있다. 석탄화력발전소는 전통적으로 대량의 대기오염물질을 배출하는 시설로 입자상 오염물질(TSP, PM_{10}, $PM_{2.5}$)뿐만 아니라 석탄 내 함유되어 있는 황성분에 의한 SOx, 연료 및 대기 중 N_2와 반응에 의해 생성되는 NOx 등 다양한 오염물질들을 배출하고 있다. 정부는 대형배출시

설에 대한 대기오염물질 배출량 조사를 위해 굴뚝에 오염물질자동전송장치(TMS, Tele-monitoring system)를 설치하여 운영 중에 있는데 2019년 기준 충청남도에서는 TMS로 산정된 대기오염물질 중 발전3사(중부, 동서, 서부)가 차지하는 비율이 66.0%로 매우 높은 비중을 차지하는 것을 나타났다. 환경부에서 매년 산정하는 대기정책지원시스템(CAPSS)에서도 충청남도의 대기오염물질 배출기여도에서 에너지·산업연소가 19.6%로 나타나 생산공정에 이어 2위로 확인되었다.



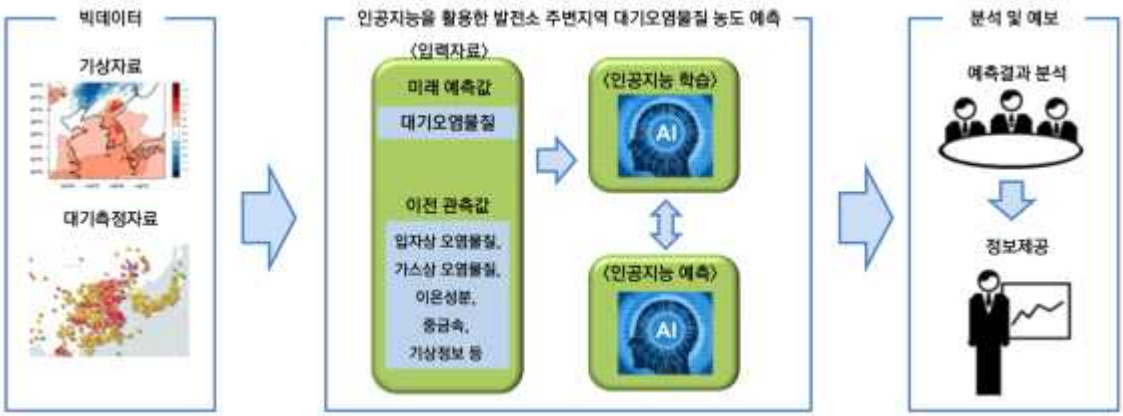
[그림 2] 국내 지역별 에너지 생산 추이 및 석탄화력발전 현황

한편, 국가에서도 이러한 대기오염물질을 모니터링하기 위하여 국가대기측정망을 운영중에 있다. 측정망 설치 위치는 목적에 따라 주거지역, 산업지역 등에 설치하고 있으나, 충청남도에는 주로 도심지역에 설치되어 있어 읍·면 단위로는 대기오염물질 정보를 상세하게 알 수 없다. 따라서, 본 연구에서는 석탄화력발전소가 위치하고 있는 당진시를 대상으로 인공지능을 활용하여 마을 단위의 미세먼지(PM_{10}) 농도를 예측하여 정보를 제공하고자 한다.

2. 연구방법

먼저, PM10 농도를 예측하기 위해서 데이터가 필요하다. 현재 충남연구원에서는 화력발전소 주변지역의 대기오염물질 정보를 수집할 수 있는 민간대기오염측정망(이하 ‘마을대기측정망’)을 운영 중에 있으며, 1시간 간격으로 데이터를 수집하고 있다. 본 데이터는 실시간으로 지역 또는 정보를 알고자 하는 모든 도민들에게 제공되고 있다.

본 연구에서는 마을대기측저망 데이터 주변 대기오염물질 농도(PM_{10}, $PM_{2.5}$, O_3, NO_2, SO_2, CO, 풍향, 풍속, 온도, 습도)를 활용하여 시간에 따른 대기오염물질 농도를 예측하고자 한다(그림 7).



[그림 3] PM10 농도 예측 프로세스

제2장

대기오염물질 예측 관련 문헌조사

1. 대기오염물질 예보시스템
2. 지하역사 미세먼지 센서 오류 감지 및 데이터 복원
3. 대기오염물질 자료 관리

제2장

대기오염물질 예측 관련 문헌조사

1. 대기오염물질 예보시스템

부산시보건환경연구원은 시민이 동네별 대기질 정보를 확인하고 대기오염에 대비할 수 있도록 ‘동네별 초미세먼지 실시간 정보’ 서비스를 제공하고 있다. 초미세먼지($PM_{2.5}$) 정보는 부산지역 27곳 대기오염 측정소 실제 관측자료와 대기오염 배출량, 토지 피복도, 인구, 기상 자료 등을 토대로 예측한 대기질 농도 시뮬레이션(대기질 진단평가시스템) 자료를 융합해 만든 것이다. 대기질 자료를 가로세로 9㎞ 범위에서 제공하는 국립환경과학원과 달리 가로세로 1㎞ 범위의 대기질 농도를 알려 줄 수 있는 장점이 있다.

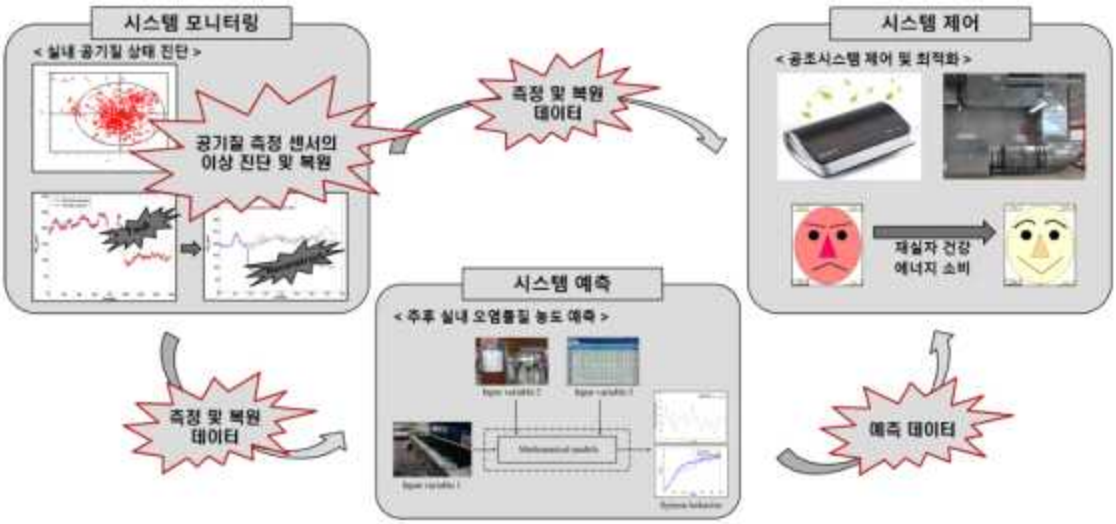
[그림 4] 부산시보건환경연구원에서 사용중인 PM2.5 예측모델

이 예측모델은 부산지역 206개 전 행정동 주민센터를 위치기반으로 예측된

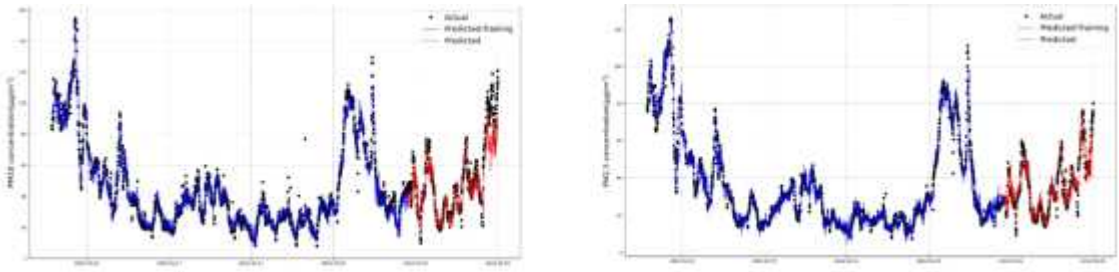
초미세먼지 정보(좋음・보통・나쁨・매우나쁨 4단계)를 연구원 홈페이지에서 제공한다. 본 예측모델은 초미세먼지 정보의 해상도를 높여 주민들에게 좀 더 가까운 대기질 정보를 제공하고 있다.

2. 지하역사 미세먼지 센서 오류 감지 및 데이터 복원

IoT센서 기술이 발전함에 따라 센서를 이용한 다양한 측정이 수행되고 있다. 센서는 원격을 활용해 실시간으로 데이터를 송・수신할 수 있는 장점이 있는 반면, 내구연한이 짧아 센서에 데이터에 대한 신뢰성 검증 주기를 짧게 가져가야 한다. 한국철도기술연구원에서는 인공지능을 활용하여 실내공간에 설치하여 사용되는 센서에 대하여 오류값을 탐지하는 연구를 수행하고 있다. 실시간으로 수집되는 데이터에 대하여 이상치를 판별하며, 이상데이터로 판별된 데이터는 복원작업을 통해 측정기기 보정이 이루어질 때까지 새로운 데이터를 생산하는 방식으로 운영하고 있다. 더 나아가 이러한 센서 데이터를 활용하여 실내공기질을 관리할 수 있는 공기청정기와 공조설비를 운영하는 연구를 진행하고 있다.



[그림 5] IoT센서 이상데이터 판별 프로세스 및 자동제어시스템 운영방식



[그림 8] 인공지능을 활용한 데이터 학습 및 예측 사례

3. 대기오염물질 자료 관리

A사에서는 관측되는 실시간 자료를 인공지능을 활용하여 품질관리하는 모델을 개발(이상감지모델)하여 운영중에 있다. 이상감지모델은 오류데이터의 1차 정제를 통해 기존에 수동으로 수행하던 자료 품질관리를 자동으로 수행함으로써 정확성과 신속성을 확보하고 있으며, 특정 시간의 관측자료가 이상 수치인지 판단하기 위해 이전 관측자료의 트렌드를 분석한다. 학습된 트렌드와 비교하여 특정 시간의 값이 정해진 범위 안에서 벗어나면 이상자료로 판단한다. 해당 모델을 사용하면 기존에 수동으로 1차 이상데이터를 판별하며 소비하는 시간을 단축할 수 있으며, 실시간으로 데이터의 통계치를 업데이트하기 때문에 정확한 이상치를 판별할 수 있는 장점이 있다.

[그림 6] A사의 이상감지모델 프로세스

제3장

화력발전소 인근 민간대기측정망 데이터 전처리 및 통계분석

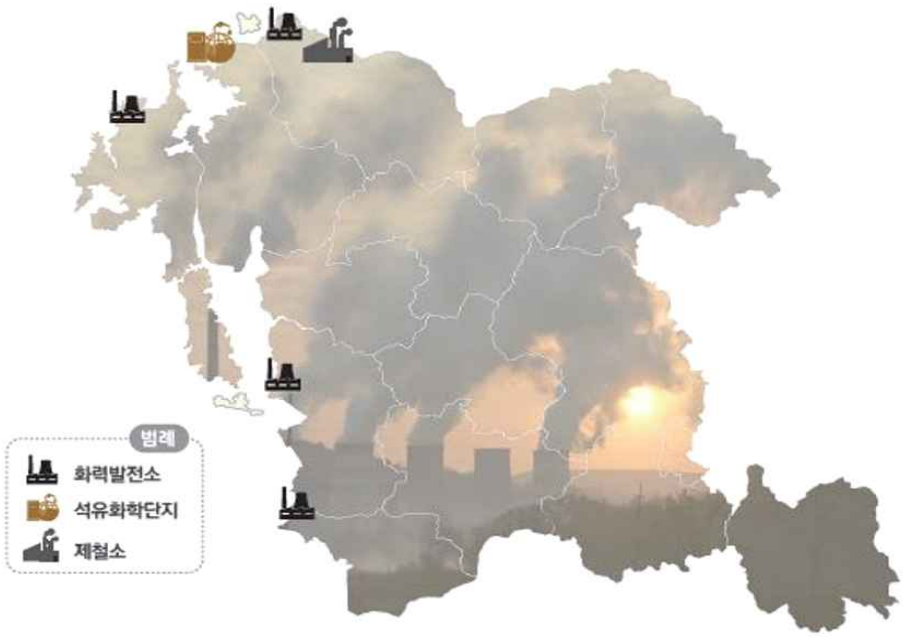
1. 마을대기측정망 개요
2. 데이터 전처리
3. 통계분석

제3장

화력발전소 인근 민간대기측정망 데이터 전처리 및 통계분석

1. 마을대기측정망 개요

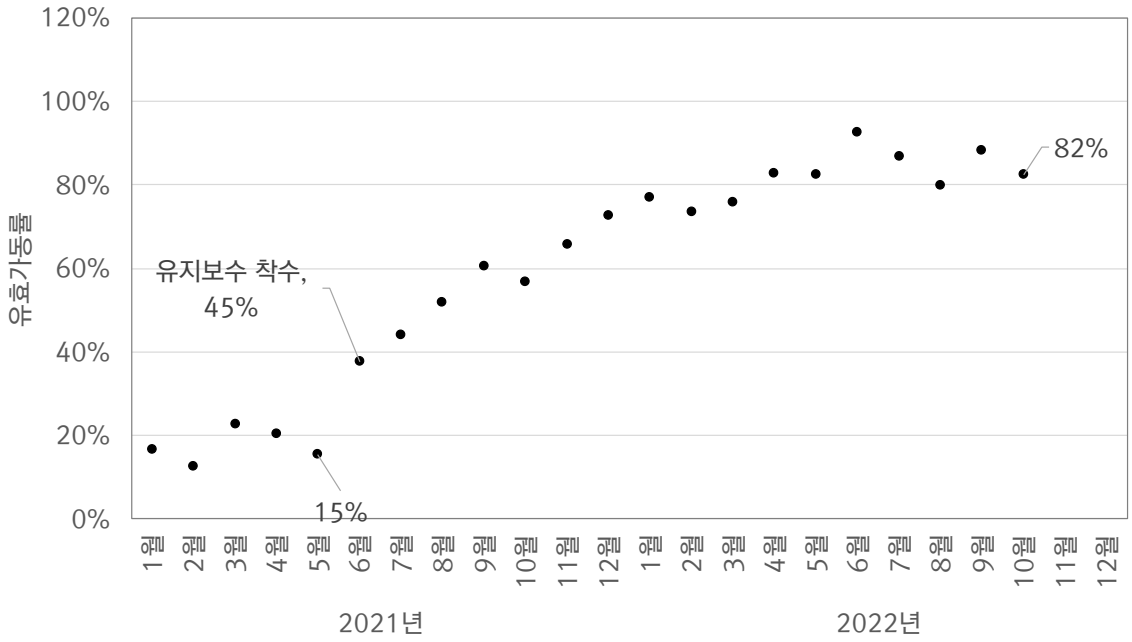
충남연구원에서는 발전소 주변 지역이 대기질 정보를 제공하기 위하여 발전소에서 운영하고 있었던 민간대기오염측정망(이하 '마을대기측정망')을 2020년부터 일괄 이관하여 통합운영관리를 수행하고 있다. 해당 사업은 '충청남도 마을대기측정망 통합정보센터 운영관리'로 명명하여 수행하고 있다.



[그림 7] 충청남도 화력발전소와 대형배출시설 위치

통합운영관리의 목적은 기존에 설치된 대기오염측정망으로부터 수집되는 데이터의 신뢰도를 향상시키고, 양질의 데이터를 제공하기 위함이다. 본 사업은 충청남도에 위치한 3개 발전사(한국동서발전, 한국중부발전, 한국서부발전)에서 각기 운영하고 있던 측정망을 2020년부터 충남연구원에서 일괄 운영하고 있다.

아래 그림에는 2021년부터 2022년까지 마을대기측정망 38개소에 대한 유효가동률 평균값을 나타내었다. 유효가동률이란 각 측정소에서 실시간으로 측정되는 대기오염물질 6개 항목 중 온전하게 사용할 수 있는 데이터의 비율을 말한다. 본 연구에서 사용된 데이터는 2022년 1월부터 2022년 11월 자료로 모두 75%이상의 유효가동률 데이터를 사용하였다.



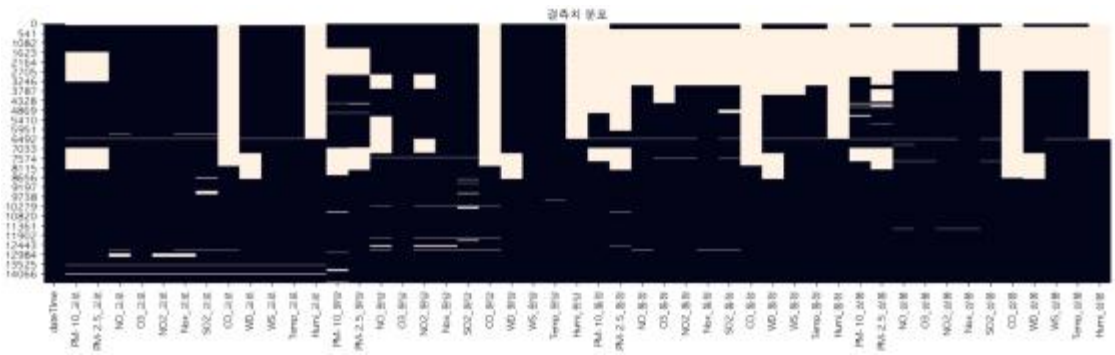
[그림 8] 2021년, 2022년 마을대기측정망 평균 유효가동률 변화

2. 데이터 전처리

실시간으로 측정되는 데이터 중 측정기기 교정, 측정기기 고장, 통신상태 문제 등 수집 데이터 중 이상자료가 발생하게 된다. 통상 이상데이터는 빈 값으로 설정하여 통계분석을 수행하게 된다. 이상데이터 판별기준은 환경부에서 고시한 대기오염측정망 설치·운영 관리지침(2021)을 참고하였으며, 아래 이상데이터 사례를 나열하였다.

- 동일값 지속
- 미세먼지/초미세먼지 역전
- 급격한 농도변화
- 기준선 농도(베이스라인) 경향 확인
- 질소산화물 농도값 비교
- 주변 측정소와의 경향 확인
- 광화학 반응물질(O_3, NO_2) 상관관계 확인

본 연구에서는 시계열 데이터를 사용하기 때문에 연속데이터가 필요하다. 이상데이터 발생에 따라 누락되어 비게 되는 데이터는 복원을 통해 연속데이터로 사용하고자 하였다. 데이터 복원 방식은 연속 5시간 미만 자료는 이동평균값을 활용하여 복원하였으며, 5시간 이상 연속으로 비는 값은 인근 측정소 값의 평균값을 적용하였다.



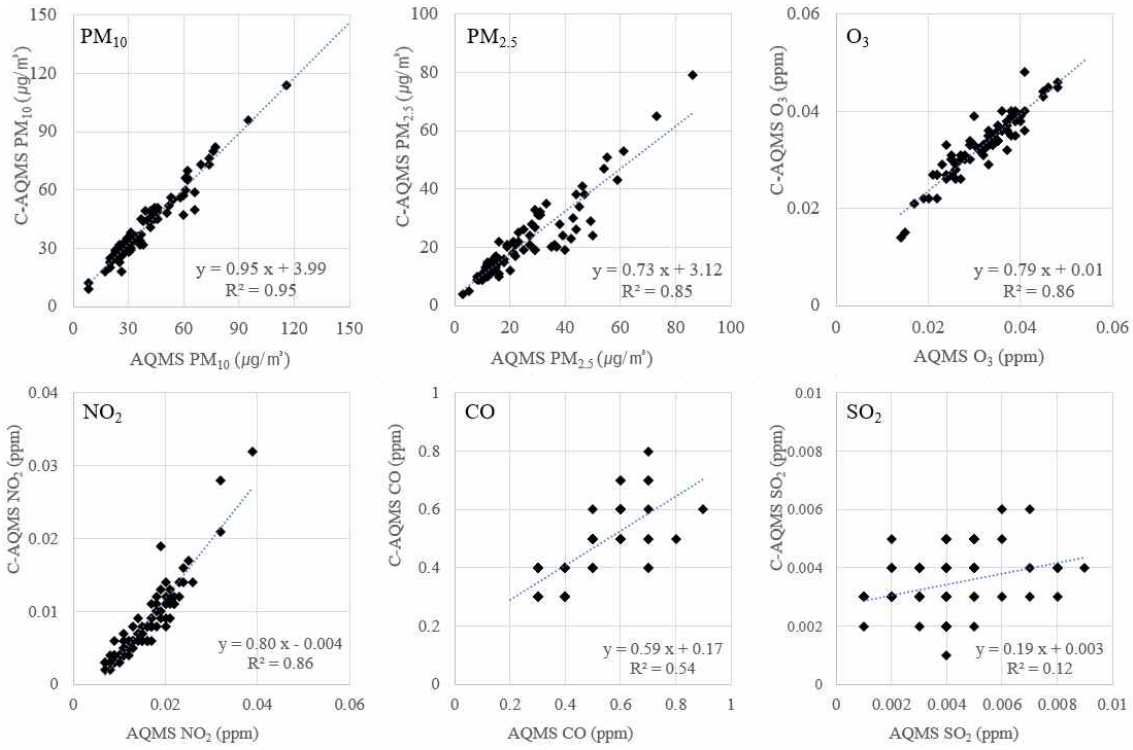
[그림 9] 결측치 분포도

3. 통계분석

1) 국가대기측정망과의 자료 비교

아래 그림에 마을대기측정망과 국가대기측정망 자료를 비교하여 산점도로 나타내었다. 측정소는 직선거리 2.6 km에 위치한 당진 중흥측정소(마을대기)와 당진시청사(국가대기)의 자료를 비교하였다.

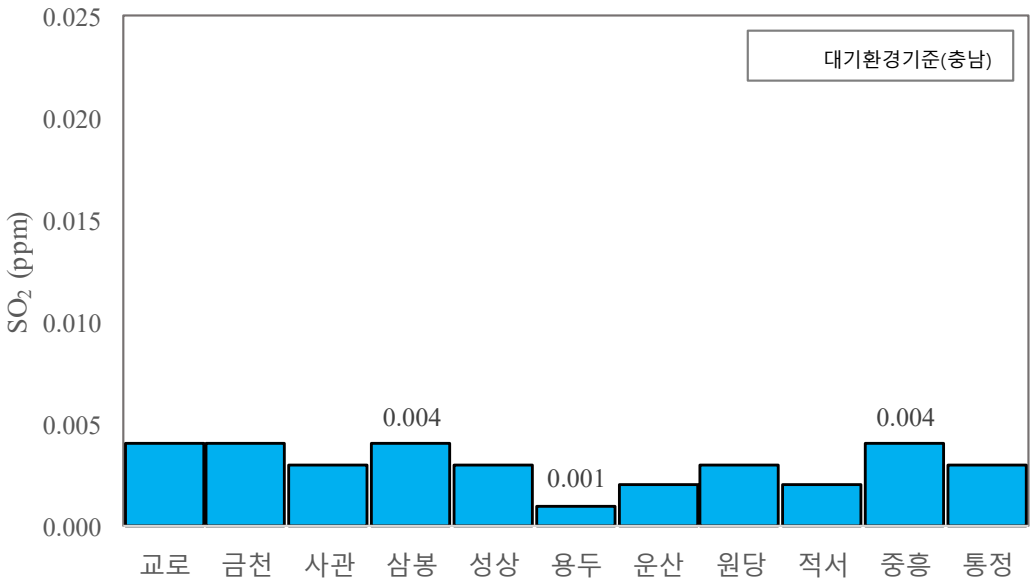
PM_{10}은 0.95로 가장 큰 상관계수(r^2, correlation coefficient)를 나타내었으며, SO_2는 0.12로 가장 작은 상관계수를 나타내었다. 가스상 오염물질(SO2, CO)의 경우, 측정값의 유효 자릿수로 인한 영향이 큰 것으로 판단된다. CO의 경우, 소숫점 첫째짜리까지만 표기하기 때문에 r^2가 낮은 것으로 판단된다. SO2의 경우, 주오염원은 연료의 연소공정에서 발생하는 것이지만, 연료 속 황 성분을 관리하기 시작하고부터 농도값 변화폭이 매우 낮은 것으로 관측되었다. 나머지 오염물질은 높은 r^2를 나타내어 마을대기측정망 자료는 국가대기측정망 수준의 데이터로 활용할 수 있다는 것을 확인하였다.



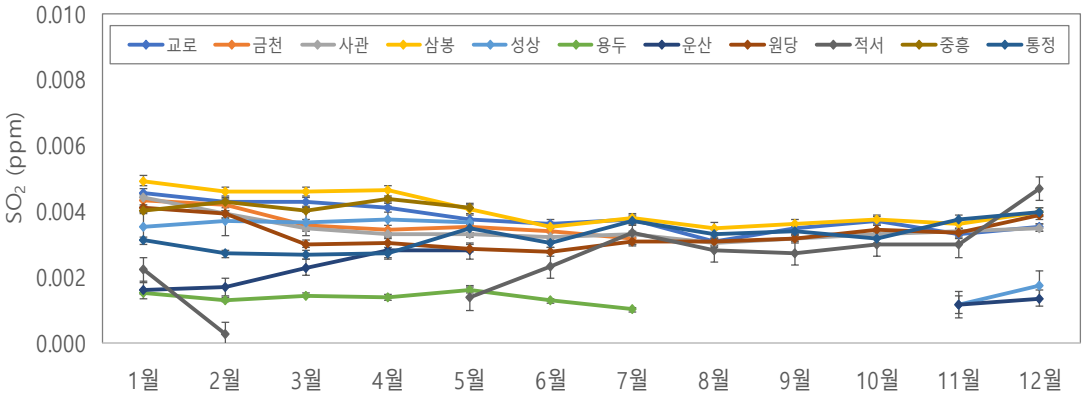
[그림 10] 국가대기측정망(당진시청사)와 마을대기측정망(중흥) 자료 산점도

2) 연평균, 월평균 대기오염물질 농도 수준

아래 그림에 당진지역 측정소별 연평균 아황산가스(SO2) 농도를 나타내었다. 충청남도 대기환경기준 중 아황산가스는 연평균 0.01 ppm으로 관리되고 있다. 당진 10개소의 연평균 SO_2는 기준치 이하로 측정되었다. 측정소별 연평균 농도분포를 살펴보면 0.001~0.004 ppm을 확인할 수 있었다. 2021년 연평균 농도(0.002~0.005 ppm)보다 약 0.001 ppm 감소한 것을 확인할 수 있었다. 11개소 측정소 중 삼봉과 중흥 측정소에서 0.004 ppm으로 가장 높은 농도를 나타내었으며, 2021년도와 같이 용두 측정소에서 0.001 ppm으로 가장 낮은 농도를 확인하였다. 아래 그림에 측정소별 월별 아황산가스(SO_2)의 연편균 농도분포를 나타내었다.

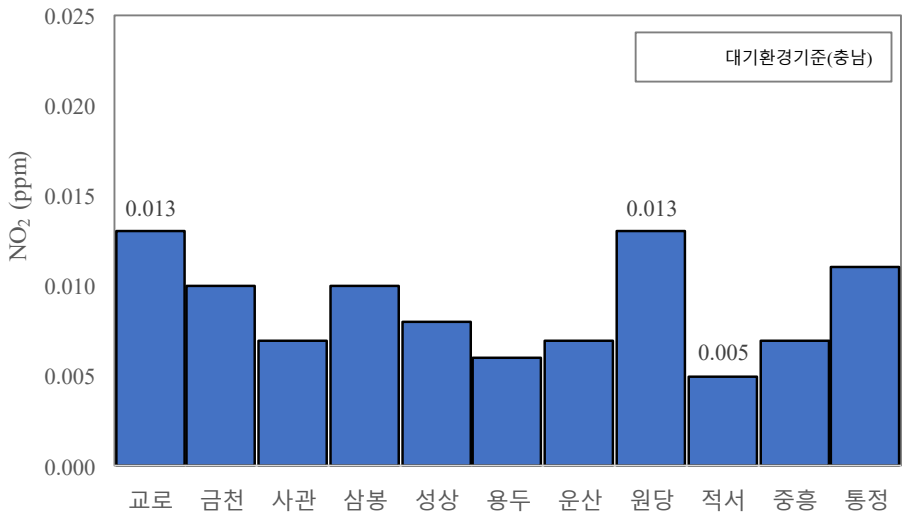


[그림 11] 당진지역 측정소별 연평균 아황산가스(SO_2) 농도>



[그림 12] 당진지역 측정소별 월별 SO_2 농도 변화>

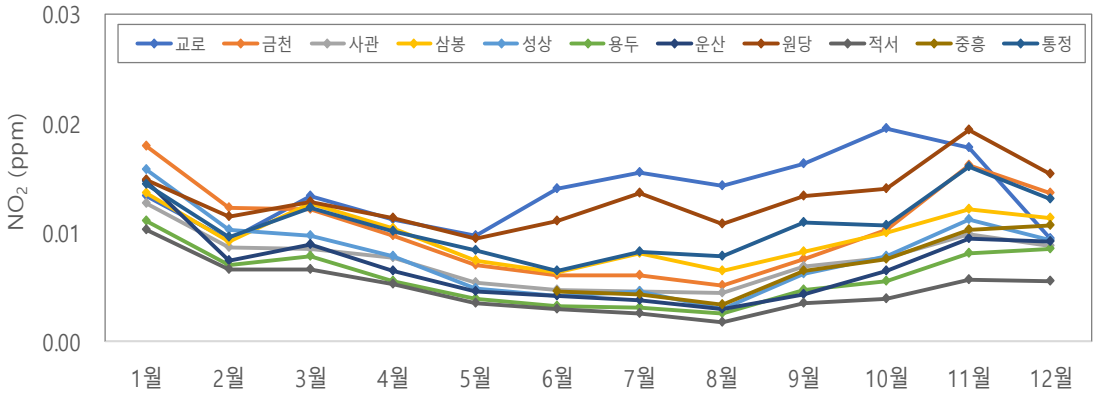
아래 그림에 당진지역 측정소별 연평균 이산화질소(NO_2) 농도를 나타내었다. 당진지역 모든 측정소에서 연간 환경기준(충청남도 NO_2 연간 기준농도는 0.02 ppm) 이하로 측정되었다. 교로와 원당 측정소는 0.013 ppm으로 가장 높은 농도를 나타내었다. 적서 측정소에서는 0.005 ppm으로 가장 낮은 농도를 나타내었다. 적서 측정소는 2021년 연평균 농도(0.006 ppm) 역시 가장 낮은 측정소로 보고한 바 있다.



[그림 13] 당진지역 측정소별 연평균 이산화질소(NO_2) 농도>

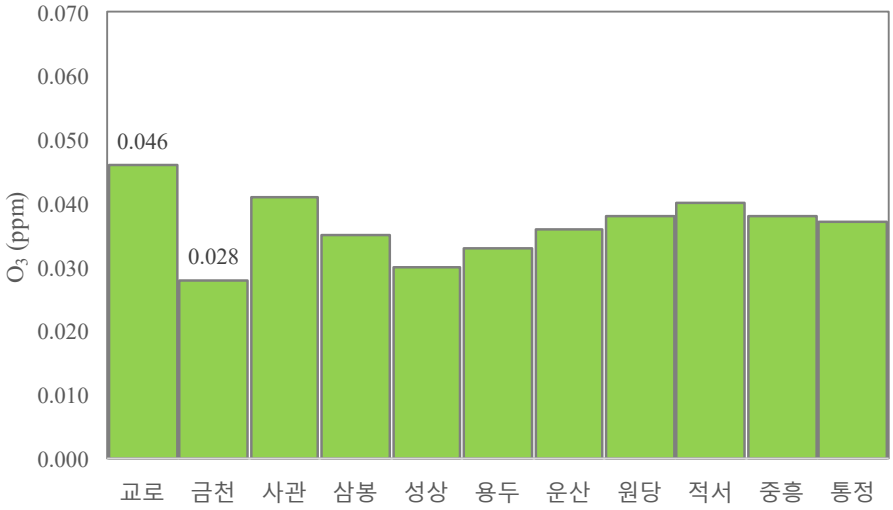
아래 그림에 당진지역 측정소별 월별 NO_2 농도 변화를 나타내었다. 발전

소와 근접한 위치에 있는 교로, 원당, 통정 측정소에서는 6월부터 농도가 증가했다가 12월에 감소하는 것을 확인할 수 있었다. 특히 11월에 교로 측정소와 원당 측정소는 연평균 기준치 농도와 유사하게 나타났다.



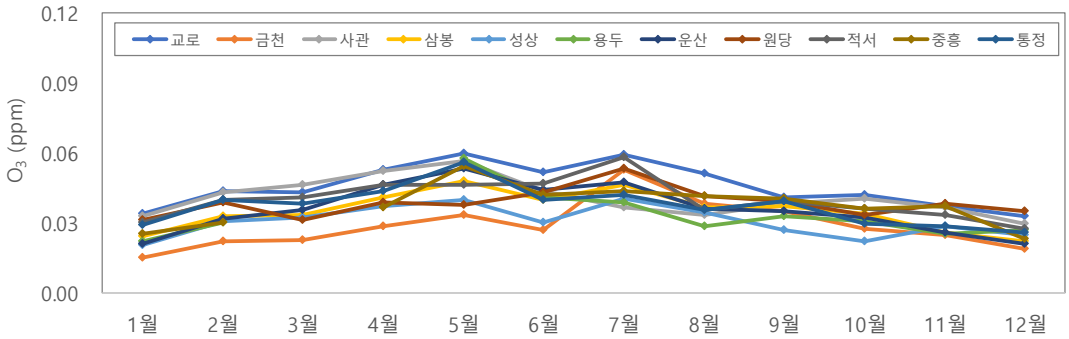
[그림 14] 당진지역 측정소별 월별 NO_2 농도 변화>

아래 그림에 당진지역 측정소별 연평균 오존(O_3) 농도를 나타내었다. 당진지역 11개소의 연평균 O_3 농도는 0.037 ppm으로 확인되었다. 2021년 0.032 ppm보다 약 0.005 ppm이 상승한 것을 확인할 수 있었다. 전국 대기오염측정망에서 수집된 오존 측정값은 매년 상승하는 것으로 보고된 바, 마을대기측정망 역시 유사한 경향을 확인할 수 있었다. 2021년과 동일하게 교로 측정소는 0.046 ppm으로 가장 높은 농도를, 금천 측정소는 0.028 ppm으로 가장 낮은 농도를 나타내었다.



[그림 15] 당진지역 측정소별 연평균 오존(O_3) 농도>

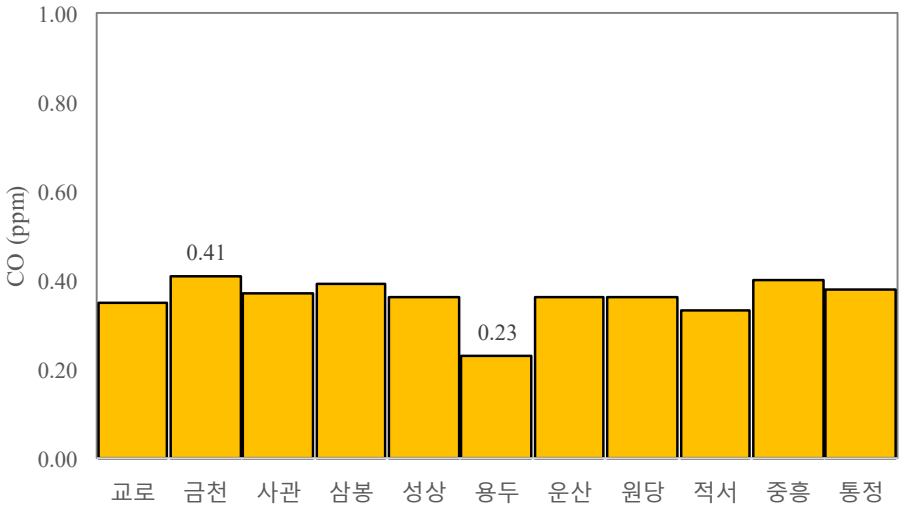
아래 그림에 당진지역 측정소별 월별 O_3 농도 변화를 나타내었다. 7월 교로 측정소에서 0.062 ppm으로 가장 높은 농도를 확인할 수 있었다. 교로 측정소는 해안과 인접한 곳에 위치하고 있어 여름철 강한 자외선으로 인하여 오존 생성이 강하게 나타난 것으로 파악되며, 다른 해안과 인접한 대기오염측정망에서도 유사한 결과를 확인하였다.



[그림 16] 당진지역 측정소별 월별 O_3 농도 변화>

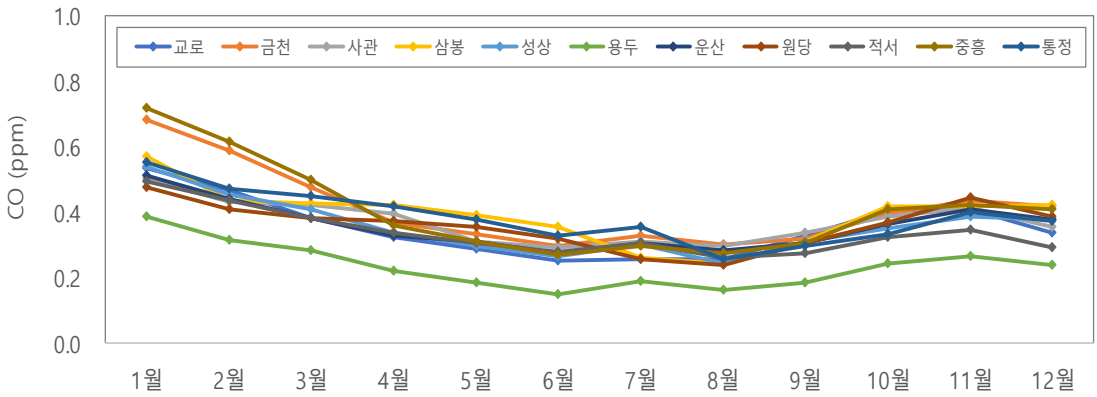
아래 그림에 당진지역 측정소별 연평균 일산화탄소(CO) 농도를 나타내었다. 당진지역 11개소의 연평균 CO 농도는 0.036 ppm으로, 2021년 연평균 농도보다 약 0.013 ppm 감소한 것으로 확인되었다. 금천 측정소는 0.41 ppm으로 가장 높은 CO 농도를 나타내었으며, 용두 측정소는 0.3 ppm으로 가장 낮은

농도를 나타내었다. 이는 2021년도 측정 결과야 동일하다.



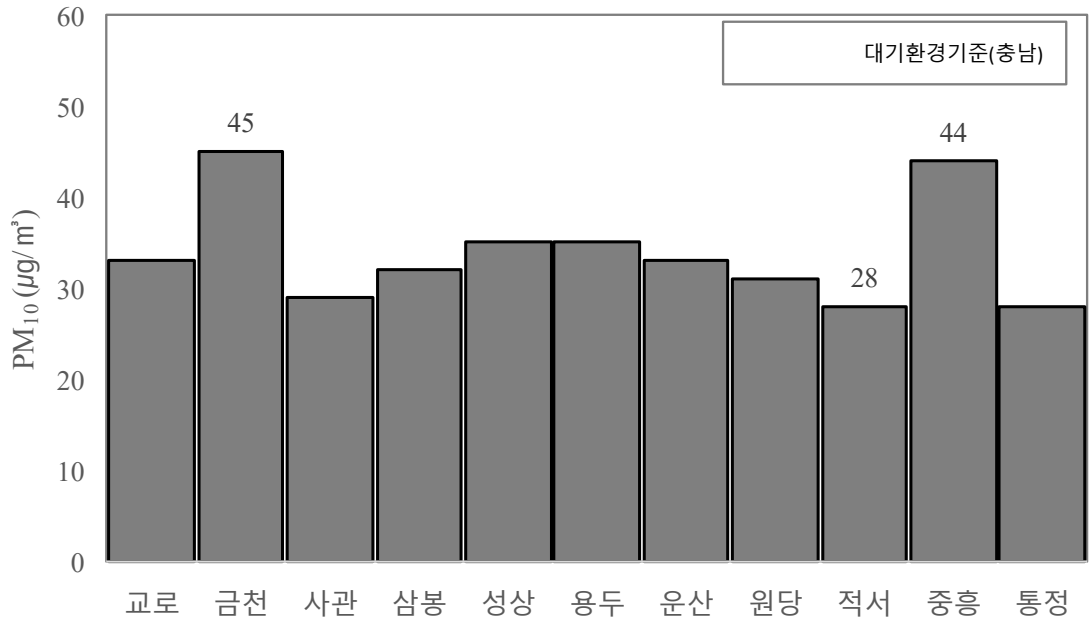
[그림 17] 당진지역 측정소별 연평균 일산화탄소(CO) 농도>

아래 그림에 당진지역 측정소별 월별 CO 농도 변화를 나타내었다.



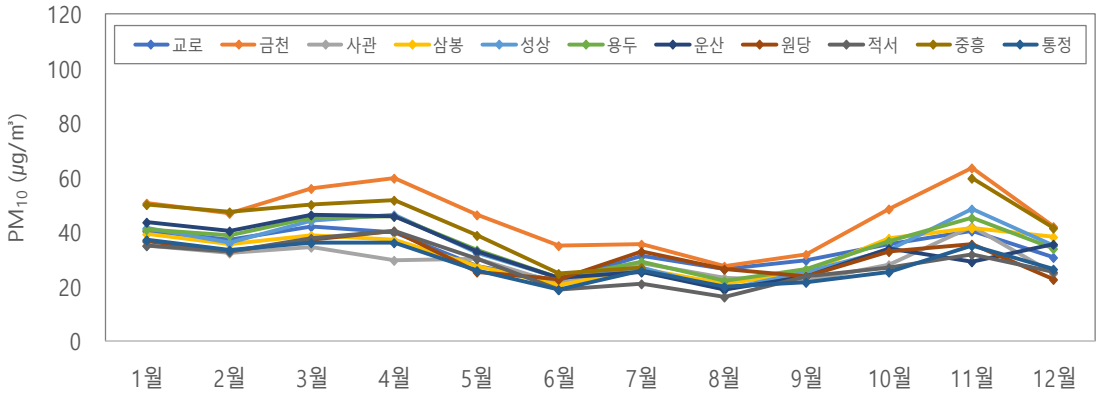
[그림 18] 당진지역 측정소별 월별 CO 농도 변화>

아래 그림에 당진지역 측정소별 연평균 미세먼지(PM10) 농도를 나타내었다. 당진지역 11개소의 연평균 PM_{10} 농도는 34 ㎍/㎥로 확인되었다. 금천 측정소에서 45 ㎍/㎥로 가장 높은 농도 값을 나타냈으며, 적서 측정소에서 28 ㎍/㎥로 가장 낮은 값을 나타내었다. 금천과 중흥 측정소는 연간 환경기준(충청남도 PM_{10} 연간 기준농도는 40 ㎍/㎥)을 초과하였으나, 금천 측정소의 경우 측정소와 600 m 거래 내에 서산-아산 간 산업도로(왕복 4차선)이 위치하고 있으며, 중흥 측정소의 경우 현대제철 당진공장과 3 km 이격거리에 위치하여 화력발전소가 아닌 다른 미세먼지 발생원의 영향이라고 판단된다.



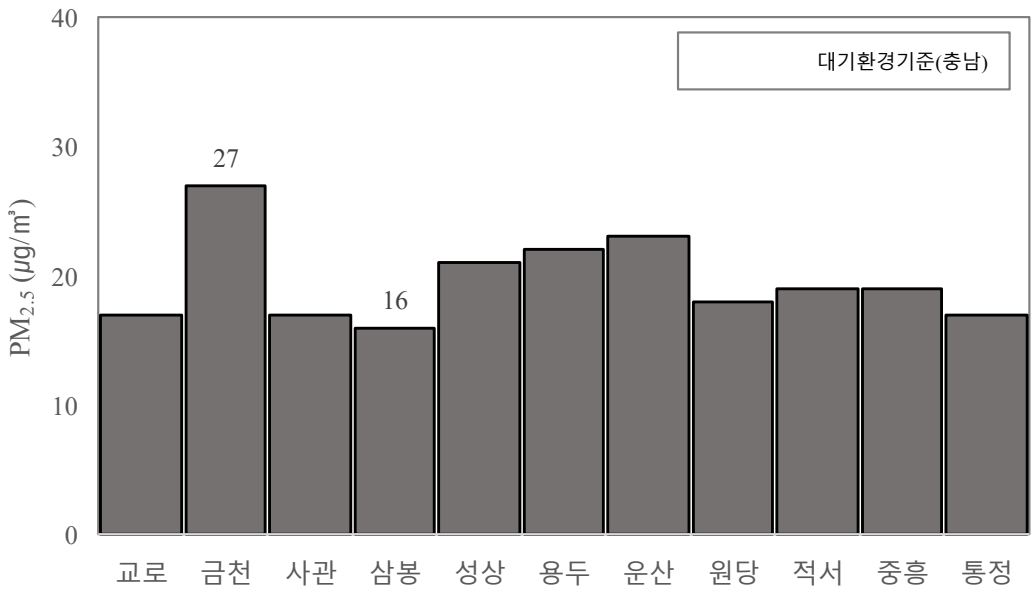
[그림 19] 당진지역 측정소별 연평균 미세먼지(PM_{10}) 농도>

아래 그림에 측정소별 월별 PM_{10} 농도 변화를 나타내었다. 월별 농도 역시 금천 측정소가 가장 높은 값을 나타냈으며, 여름철에 농도가 제일 낮고 겨울과 봄철에 농도가 증가하는 것을 확인할 수 있었다.



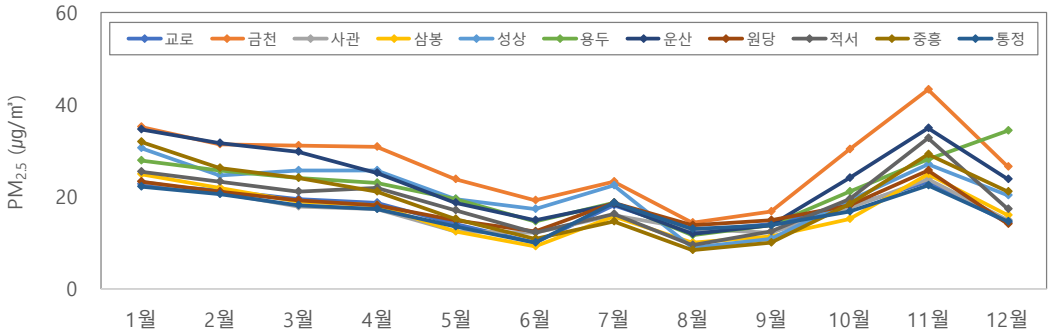
[그림 20] 당진지역 측정소별 월별 PM_{10} 농도 변화>

아래 그림에 당진지역 측정소별 연평균 초미세먼지($PM_{2.5}$) 농도를 나타내었다. 당진지역 11개소의 연평균 $PM_{2.5}$ 농도는 20 ㎍/㎥로 4개 시·군 중 가장 높은 농도로 확인되었다. 11개 측정소 모두 연간 환경기준(충청남도 $PM_{2.5}$ 연간 기준농도는 15 ㎍/㎥)을 초과하였으며, 금천 측정소는 27 ㎍/㎥로 가장 높은 초미세먼지 농도를 나타내었다. 연평균 농도가 가장 낮은 측정소는 삼봉 측정소로 확인되었으며, 연평균 농도는 16 ㎍/㎥를 나타내었다.



[그림 21] 당진지역 측정소별 연평균 초미세먼지($PM_{2.5}$) 농도>

아래 그림에 측정소별 월별 초미세먼지 농도 변화를 나타내었다. 가을철(11월)에 가장 높은 농도를 확인할 수 있었다.



[그림 22] 당진지역 측정소별 월별 $PM_{2.5}$ 농도 변화>

제4장

측정소 미설치지역의 대기오염물질 농도 예측

1. 적용한 인공지능 기법 상세
2. 대기오염물질 예측 결과

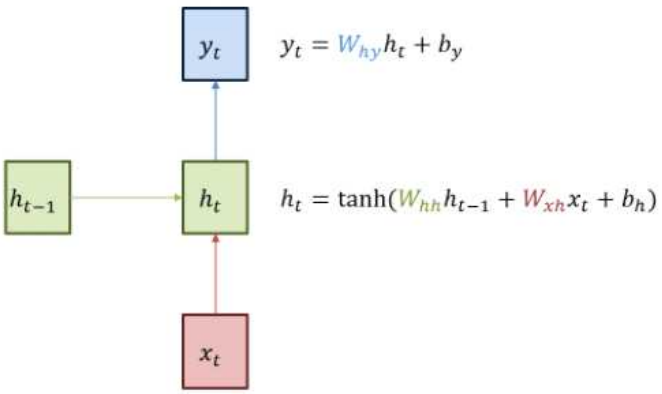
제4장

측정소 미설치지역의 대기오염물질 농도 예측

1. 적용한 인공지능 기법 상세

1) 순환신경망(RNN, Recurruent Neural Network)

기존 머신러닝 기법과 순환신경망(Recurrent Neural Network;RNN)의 가장 큰 차이점은 기존 데이터를 활용하여 가중치(weight)를 스스로 업데이트 하는 것이다. RNN은 히든 노드가 방향을 가진 엣지로 연결돼 순환구조를 이루는(directed cycle) 인공신경망의 한 종류이다. 음성, 문자 등 순차적으로 등장하는 데이터 처리에 적합한 모델로 알려져 있으며, Convolutional Neural Networks(CNN)과 더불어 최근 들어 주목받고 있는 알고리즘이다<Fig. 2.8-9>. RNN은 시퀀스 길이에 관계없이 input과 output을 받아들일 수 있는 네트워크 구조이기 때문에 필요에 따라 다양하고 유연하게 구조를 만들 수 있다는 점이 RNN의 가장 큰 장점이다.



[그림 23] RNN의 기본원리

2) 장단기메모리(LSTM, long-short term memory)

단기메모리(Long short term memory;LSTM)는 여러 종류의 게이트가 있어 입력을 선별적으로 허용하고, 계산 결과를 선별적으로 출력할 수 있다. LSTM의 핵심은 열고 닫는 것이며, 게이트는 0에서 1사이의 실수값을 가지고 개폐 정도를 조절한다. 언제 얼마만큼 여닫을지는 학습으로 알아낸다. 구조는 RNN의 은닉층에 메모리블록을 배치한 것이다. 이 메모리 블록에는 입력 게이트와 출력 게이트가 있다.

최종 En+2를 구하는 방법은 Fig.에 나타내었다. 최종 weight 업데이트 방법은 (식 1-3)에 나타내었다. 각 hidden state에서 계산되는 방법을 총 더하면 최종 weight 업데이트 양이 구해진다.

$$W = W - larning\_rate \bullet \frac{\partial E}{\partial W} \qquad \text{식 (1)}$$

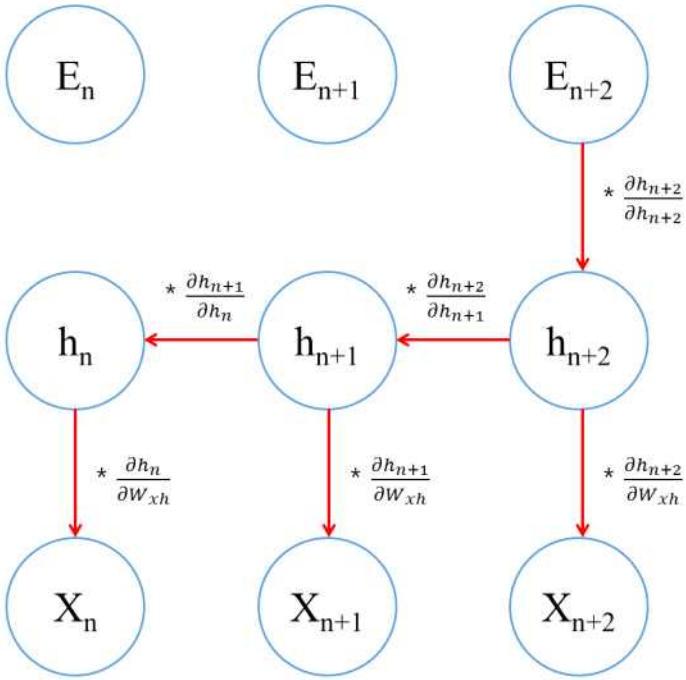
$$\frac{\partial E_{n+2}}{\partial W} = \frac{\partial E_{n+2}}{\partial h_{n+2}} \times \frac{\partial h_{n+2}}{\partial W_{xh}} + \frac{\partial E_{n+2}}{\partial h_{n+2}} \times \frac{\partial h_{n+2}}{\partial h_{n+1}} \times \frac{\partial h_{n+1}}{\partial W_{xh}} + \frac{\partial E_{n+1}}{\partial h_{n+1}} \times \frac{\partial h_{n+1}}{\partial h_n} \times \frac{\partial h_n}{\partial W_{xh}} \qquad \text{식 (2)}$$

$$\frac{\partial E}{\partial W} = \sum_{i=1}^{n} \frac{\partial E_n}{\partial W_n} \qquad \text{식(3)}$$

$$\frac{\partial E_{n+2}}{\partial W} = \frac{\partial E_{n+2}}{\partial h_{n+2}} \times \frac{\partial h_{n+2}}{\partial W_{xh}} + \frac{\partial E_{n+2}}{\partial h_{n+2}} \times \frac{\partial h_{n+2}}{\partial h_{n+1}} \times \frac{\partial h_{n+1}}{\partial W_{xh}} + \frac{\partial E_{n+1}}{\partial h_{n+1}} \times \frac{\partial h_{n+1}}{\partial h_n} \times \frac{\partial h_n}{\partial W_{xh}} \qquad \text{식 (4)}$$

hidden state에서는 데이터의 양이 많아질 경우, 데이터 수와 비례하여 weight update가 수행이 되어진다. learning rate와 곱해지는 미분값이 연속적으로 곱해질 경우, 미분값이 1보다 작은 값이면 weight는 0에 수렴하게 된다. weight를 업데이트 하는 과정에서 기존 weight와 거의 차이가 없어지게 되는 현상이 발생하는데, 이와 같은 현상을 그라디언트 베니싱(Gradient Vanishing)이라고 한다. 그라디언트 베니싱이 발생하게 되면 weight update가 비효율적이 되고, 반대로, 미분값이 1보다 클 경우, 여러번 반복될 경우 무한대로 발산하는 오류가 발생하게 된다. 이를 그라디언트 익스플로딩(Gradient Exploding)이라고 한다. 이와 같은 단점들을 보완하기 위하여 만들어진 기법이

LSTM이며, LSTM에 대한 개요도를 아래 <그림 24-25>에 나타내었다. LSTM의 첫 번째 절차는 모든 데이터를 학습하는 것이 아니라, 일부 데이터만을 가지고 학습을 진행하는 것이다. 기존 데이터(ct-1)가 LSTM cell로 입력이 되고, 새로 입력되는 x값이 기존 hidden state(ht-1)과 만나 sigmoid함수에 연산이 됩니다. sigmoid함수는 0에서 1사이의 값을 나타내는 함수로, 확률값이 된다. 즉 입력되는 전체 값의 일부만 기억하고 나머지 데이터는 두 버리는 방식이다.



[그림 24] Schematic diagram of backpropagation by chain-rule

다음은 새로 들어온 데이터에 대한 학습인지 단계입니다. 일부 데이터를 버리고 나머지 데이터를 학습시키는 절차가 진행된다. 기존 hidden state와 새로 입력되는 일부 데이터가 sigmoid 함수와 tangent sigmoid 함수에 곱해진다. 이 과정을 통하여 기존에 학습된 데이터에 새로 들어온 데이터가 메모리셀(Memory cell)에 추가된다.

마지막 절차는 메모리셀에 저장된 데이터가 tanh를 통하여 hidden state로 입력되고, hidden state에 있는 현재의 정보의 일부가 sigmoid함수를 통하여 더해져 hidden state에 저장이 된다. 저장된 값은 output 값으로 출력이 됨과

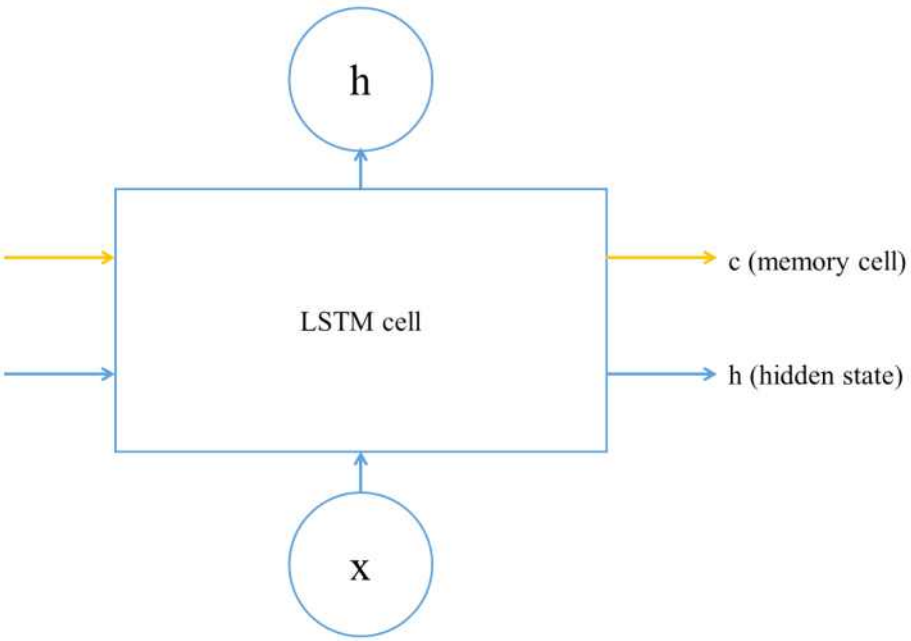
동시에, 다음 hidden state로 이동이 되어 진다. 이와 같은 절차를 아래 (식 5~8)에 나타내었다.

$$i_t = \sigma(W_{xh\_i}x_t + W_{hh\_i}h_{t-1} + b_{h\_i}) \quad \text{식 (5)}$$

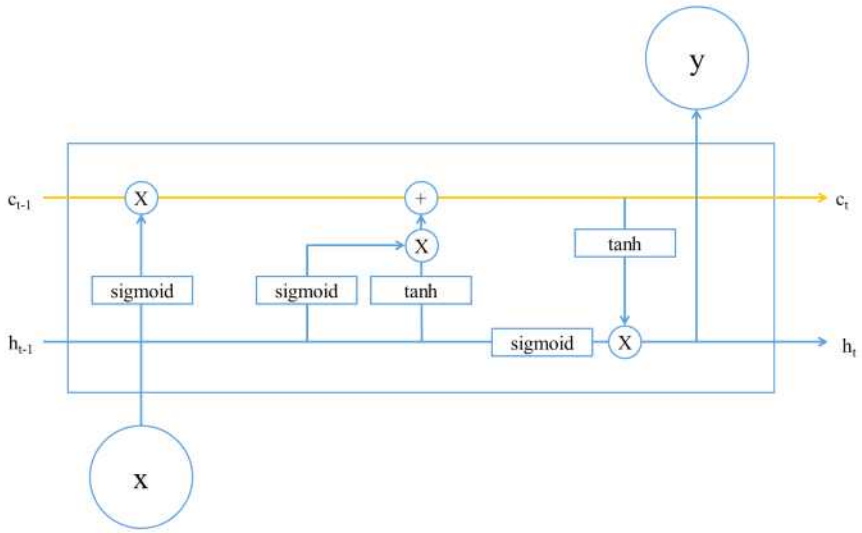
$$o_t = \sigma(W_{xh\_o}x_t + W_{hh\_o}h_{t-1} + b_{h\_o}) \quad \text{식 (6)}$$

$$g_t = \tanh(W_{xh\_g}x_t + W_{hh\_g}h_{t-1} + b_{h\_g}) \quad \text{식 (7)}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad \text{식 (8)}$$



[그림 25] LSTM basic structure



[그림 26] Unit cell structure of Long-Short Term Memory

2. 예측 대상지점 선정 및 데이터 수집

미래 대기오염물질 예측을 위하여 당진과 보령을 선정하였으며, 선정된 지점에 위치한 측정망은 발전소로부터 10 km 이내에 위치하고 있는 측정소 4개를 선정하였다. 당진지역은 교로, 원당, 삼봉, 통정 측정소의 데이터를 활용하였다. 보령지역은 오포, 송학, 주포, 죽정 측정소를 선정하였다. 예측 대상지점은 각 지역의 4개 측정소 중 가운데 위치해 있는 지점을 선정하였다. 당진은 원당 측정소를, 보령은 송학 측정소를 예측 대상 지점으로 선정하였으며, 선정 지점의 미래 후 대기오염물질 농도를 예측하였다.

마을대기측정망에는 총 12개 항목의 데이터(PM10, PM2.5, NO, NO2, NOx, SO2, CO, O3, 온도, 습도, 풍향, 풍속)가 실시간으로 수집된다. 데이터 수집은 2022년 1월 1일부터 2022년 11월 31일까지 데이터를 활용하였으며, 1시간 간격의 대기오염물질 및 기상자료를 활용하였다. 본 연구에서는 예측을 위하여 3개 측정소의 12개 항목 모두 입력변수로 사용되었으며, 향후 연구를 통해 예측 대상 물질에 대한 입력변수의 민감도 분석을 수행예정이다. 예측 대상물질은 PM10, PM2.5, NO2, SO2, O3, CO로 선정하였다. 위 6개 물질은 현재 대기오염측정망에서 제공하고 있는 6가지 항목이다.

대기오염측정망 데이터는 이상데이터 선별을 통해 자료가 공개되며, 이상데이터는 공란으로 표기되어 제공된다. 대기오염물질 데이터는 시계열 데이터로, 각 항목마다 이전 시간 데이터가 영향을 준다. 또한, 시계열 자료를 입력값으로 예측을 수행할 경우, 공란이나 아웃라이어 데이터는 반드시 전처리가 필요하다. 본 연구에서도 이와 같은 공란이나 아웃라이어에 대한 전처리를 수행하였다. 공란 데이터는 24시간 평균 값으로 대체하였으며, 아웃라이어 데이터는 각 항목별로 95% 범위를 초과 값을 기준으로 제거하였다. 제거한 값은 공란 데이터로 간주하여 24시간 평균 값으로 처리하였다.

전처리된 데이터는 각 항목이 입력변수로 활용되며, 총 36개의 입력변수를 생성하였다. 데이터의 80%는 데이터 모델을 생성하는데 활용하였으며, 나머지 20%는 예측모델의 성능을 확인하기 위하여 사용되었다. 예측 결과는 나머지 20%에 대한 입력값과 예측값의 비교를 통해 예측 성능을 확인하였다.

입력변수와 예측 대상 항목을 선정하여 예측을 수행한 후, 예측 성능을 평가하기 위하여 실측값과 예측값의 평균제곱근차와 상관계수를 활용하였다.

평균 제곱근 편차(Root Mean Square Deviation; RMSD) 또는 평균 제곱근

오차(Root Mean Square Error; RMSE)는 추정 값 또는 모델이 예측한 값과 실제 환경에서 관찰되는 값의 차이를 다룰 때 흔히 사용하는 측도이다. 정밀도(precision)를 표현하는데 적합하다. 각각의 차이값은 잔차(residual)라고도 하며, 평균 제곱근 편차는 잔차들을 하나의 측도로 종합할 때 사용된다(Park et al., 2018).

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y_i} - y_i)^2}{n}} \qquad \text{식 (9)}$$

두 변수간에 어떤 선형적 또는 비선형적 관계를 가지는지를 분석하는 방법이다. 두 변수는 서로 독립적인 관계이거나 상관된 관계일 수 있으며 이때 두 변수 간의 관계의 강도를 상관관계(Correlation, Correlation coefficient)라 한다. 상관분석에서는 상관관계의 정도를 나타내는 단위로 모상관계수로 ρ를 사용하며 표본 상관계수로 r2을 사용한다..

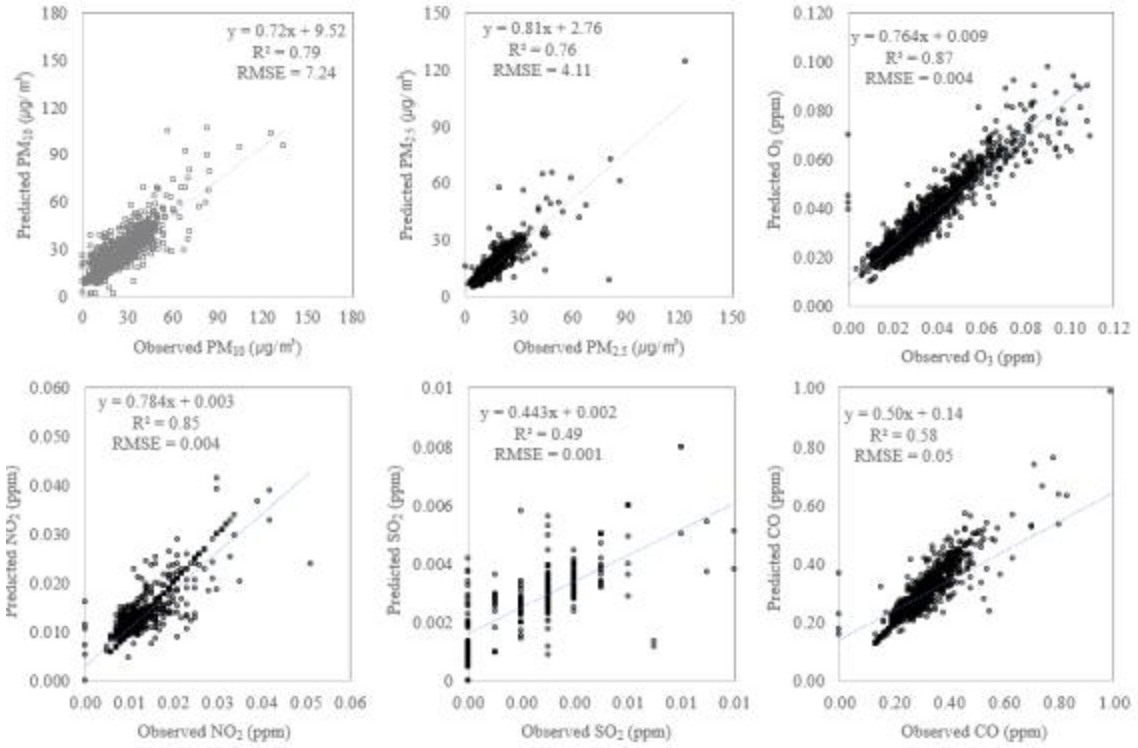
$$r^2 = \left(\frac{n(\sum_{i=1}^{n} y_i\hat{y_i}) - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} \hat{y_i})}{\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2 n\sum_{i=1}^{n} \hat{y}_i^2 - (\sum_{i=1}^{n} \hat{y}_i)^2}}\right)^2 \qquad \text{식 (10)}$$

위 식 (10)에서 n은 데이터의 개수를 나타내며, y_i는 실측값을 의미하며, $\hat{y_i}$는 예측값을 의미한다.

2. 대기오염물질 예측 결과

1시간 후 대기오염물질 농도 예측 결과를 산점도를 아래 그림 ###에 나타내었다. 그래프의 x축에는 실측값을 나타내었으며, y축에는 예측값을 나타내었다.

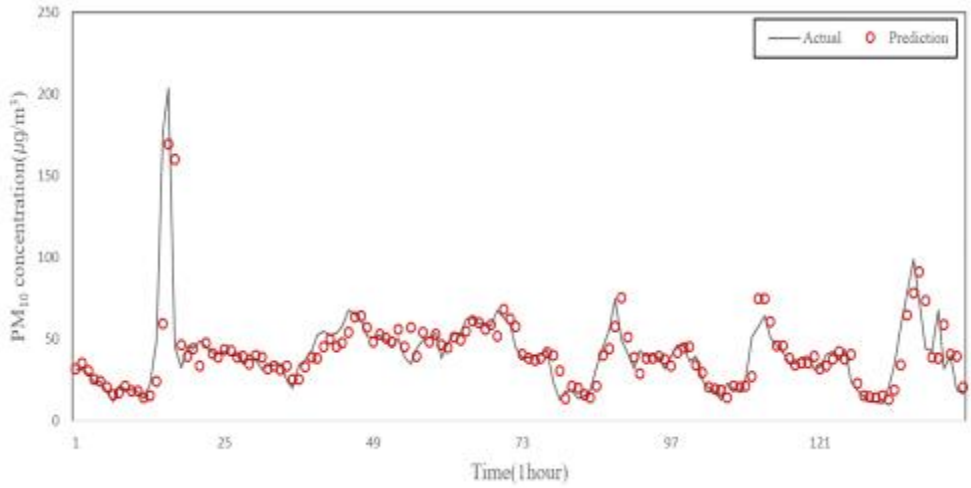
PM10과 PM2.5는 각각 0.79와 0.76의 상관계수를 확인할 수 있었으며, 7.24와 4.11의 RMSE를 확인할 수 있었다. 입자상물질은 고농도 이벤트가 발생하더라도 예측값값과 실측값의 차이가 적은 것으로 확인되었다(그림 ##). 가스상물질은 0.49~0.87의 상관계수와 0.001~0.05의 RMSE를 확인할 수 있었다. RMSE는 실측값과 예측값의 평균 제곱근 차이를 나타내는 지표로, 각 항목별로 절대값이 다르기 때문에 RMSE 절대값으로 예측 성능을 판단하기엔 다소 무리가 있다. CO값은 다른 4개 가스상오염물질보다 통상 10배이상 높은 농도를 나타내기 때문에 CO의 예측결과가 가장 나쁘다고는 볼 수 없다.



[그림 27] 당진지역 측정소 대기오염물질 농도 예측 결과

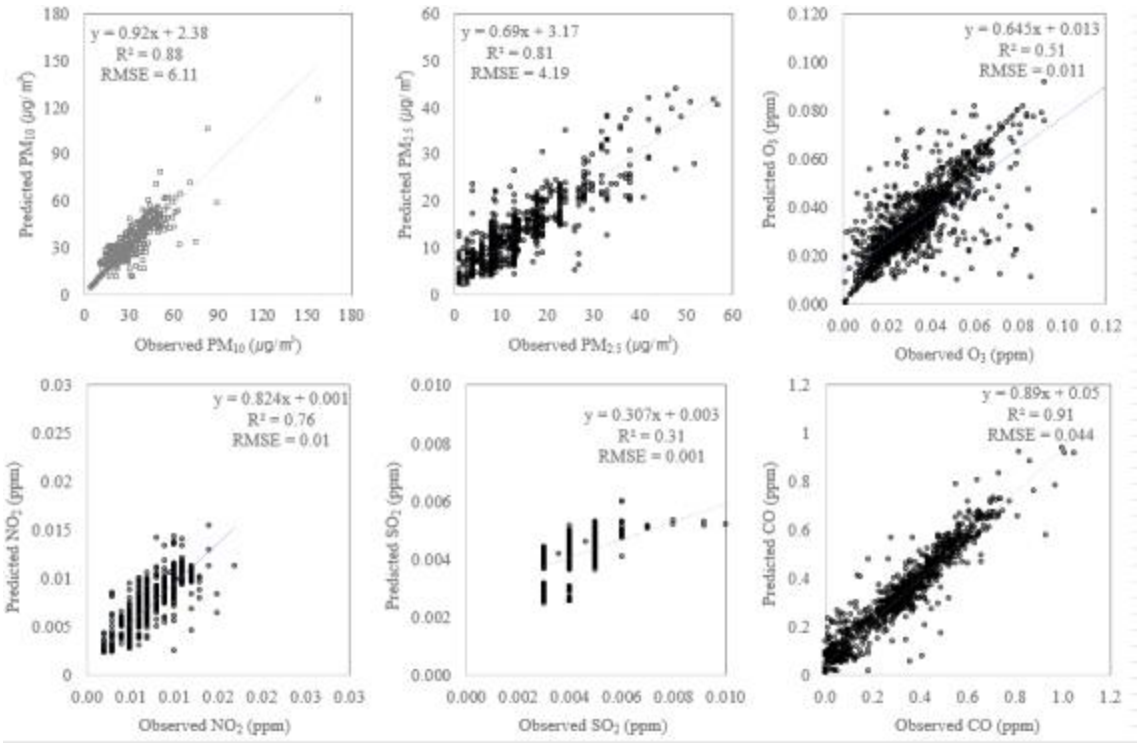
SO_2의 경우, 데이터가 0.001~0.008에 분포하고 있다. SO2의 특성상 일반 대기질에서 높은 농도로 존재하지 않고 연료의 연소 등의 공정에서 높은 농도

로 배출된다. SO2의 예측 성능이 낮게 나온 이유는 데이터의 분포도의 차이로 볼 수 있다.



[그림 28] 미세먼지 고농도 이벤트 발생시 실제값과 예측값의 비교

아래 그림에 24시간 후 대기오염물질 예측 결과를 나타내었다.



[그림 29] 24시간 후 대기오염물질 농도 예측 결과

3. 활용방안

1) 제안화력발전소 주변지역 대기오염물질 고해상도 자료 확보

본 연구를 통해 석탄화력발전소 주변지역에 대하여 측정기기 없이 대기오염물질 자료구축에 대한 가능성을 확인할 수 있었다. 현재 충남연구원에서 운영하고 있는 마을대기측정망의 경우, 각 읍·면에 위치한 행적복지센터 위에 측정소가 위치하고 있는 실정이기 때문에 리 단위로는 대기오염물질 정보를 확인할 수 없는 실정이다. 본 연구를 바탕으로 추후 마을단위의 대기오염물질 정보를 구축한다면 발전소 인근 주만들에게 정보제공 차원에서 좋은 자료가 될 것으로 판단된다.

2) 대기질 미래정보 수집에 따른 대기환경 정책수립

본 연구에서는 동시간 대에 대기오염물질 농도 산출에 주력하였으나, 시계열 데이터 기반이기 때문에 주변 측정소의 자료를 활용한다면 미래 대기오염물질 농도도 예측할 수 있다. 대기오염물질 예보는 보건환경연구원에서 수행하고 있으나, 현재 모델링 기반으로 활용하고 있기 때문에 기술력과 구현 시간에 대한 제약이 많은 실정이다. 인공지능을 활용한다면 예보 관련 시간 단축 뿐만 아니라 국소지역에 대한 농도 예보도 가능할 것으로 판단된다.

3) 대기오염측정망 설치에 따른 설치 적합성 활용

현재 매년 국가대기측정망 개수가 늘어나고 있는 추세이다. 2016년 충청남도에 단 6개의 측정망이 존재하였으나 매년 증가하여 현재 42개(도시대기, 도로변대기) 측정망이 존재하고 있다. 본 연구결과를 바탕으로 예측성능이 낮은 측정망에 대하여 측정소를 우선 적용 검토가 가능할 것으로 판단된다.

참고문헌

1. 환경공단, 굴뚝자동측정기기(TMS) 측정결과 공개(https://cleansys.or.kr)
2. 전력거래소, 전력통계정보시스템(http://epsis.kpx.or.kr)
3. 환경부, 국가대기오염물질 배출량 서비스(http://airemiss.nier.go.kr)
4. 환경부, 2019, 대기오염측정망 설치·운영 지침
5. 환경부, 2019, 미세먼지 관리 종합계획(2020~2024)
6. 환경공단, AirKorea(https://www.airkorea.or.kr)
7. 환경부, 2020, 환경정책 기본법 시행령 별표 1. 환경기준
8. 환경부, 2020, 대기환경월보 5월호
9. 김종범, 윤수향, 이상신, 김경환, 노수진, 배귀남 (2020) 충남지역 PM_{10}과 $PM_{2.5}$ 농도의 시공간 분포 특징, 한국대기환경학회 36(4), 464-481
10. 국립환경과학원, 2019, 2018 대기환경연보

연구책임	박세찬 기후변화대응연구센터 초빙책임연구원
연구참여	이상신 기후변화대응연구센터 연구위원
	김종범 기후변화대응연구센터 책임연구원
	최영남 기후변화대응연구센터 책임연구원
	이가혜 기후변화대응연구센터 연구원
	송혜영 기후변화대응연구센터 연구원

정책지원과제 2022-06

인공지능기법을 활용한 측정소 미설치 지역 PM10 농도 예측

발행일 : 2022년 12월

발행인 : 충남연구원장

발행처 : 충남연구원 서해안기후환경연구소

(32258) 충청남도 홍성군 홍북읍 홍예로 360

홈페이지 http://www.shari.re.kr

발간등록번호 : -