머신러닝을 활용한 대기환경

빅데이터 분석 및 예측

2010104050

박재호

요약

본 프로젝트는 인체에 유해한 성분의 물질이 다수 포함된 대기오염물질 데이터분석을 실시한다. 분석 범위는 경기도이며(31개 도시) 오염물질의 경향과 패턴을 세부적인 시간별, 지역별로 나누어 분석한다. 또한 분석한 데이터를 토대로 머신러닝을 실시하여 미래의 오염 추이를 예측한다.

1. 서론

1.1 연구배경

인류는 산업혁명이래로 스모그, 일산화탄소 등 끊임없는 대기오염 문제를 가지고 있었지만 2016년 세계적 대기오염의 화두는 미세먼지에 의한 대기오염이었다. 미세먼지는 인체에 유해한 성분의 물질이 다수 포함된 인위적인 대기오염물질이다. 많은 양의 미세먼지에 지속적으로 노출될 경우 인체에 치명적인 질병을 유발할 수 있을 뿐만 아니라 자연 생태계에도 큰 영향을 준다. 이러한 피해를 최소화 하기 위해서는 미세먼지 측정 및 경보 인프라 구축이 필수적이다. 본 프로젝트는 경기도권의 대기오염 경보 및 인프라 구축에 도움을 주고자 경기도권의 2014 ~ 2016년 대기오염물질을 지역별, 연도별로 분석하여 유의미한 결과를 시각화 및 머신러닝을 통한 예측을 실시하고자 한다..

1.2 연구목표

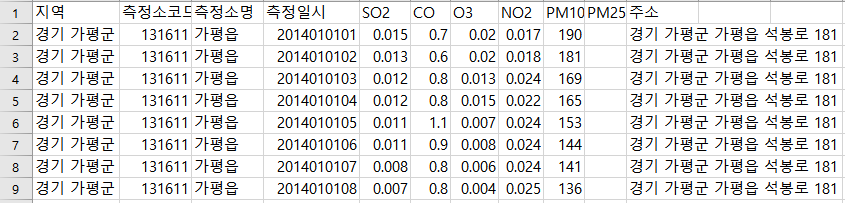
본 프로젝트의 범위는 경기도 지역으로 한정하며 다음 3가지를 목표로 한다.

1. 2014 ~ 2016년 대기오염물질 및 기상데이터를 기반으로 경기도의 31개 도시를 범위로 지역별, 시간별(년, 월, 주, 시간) 대기 오염물질 변화 추이를 분석한다. 이를 토대로 유의미한 결과를 도출하여 시각화한다..
2. 각 지역의 인구밀도와 대기오염물질의 지역특성, 공간분포 상 취약지를 고려하여 추가 대기환경 측정소 위치를 선정한다.
3. 다양한 머신 러닝 방법을 이용하여 가장 성능이 좋은 모델을 선출 후 미래 대기오염 추이를 예측한다.

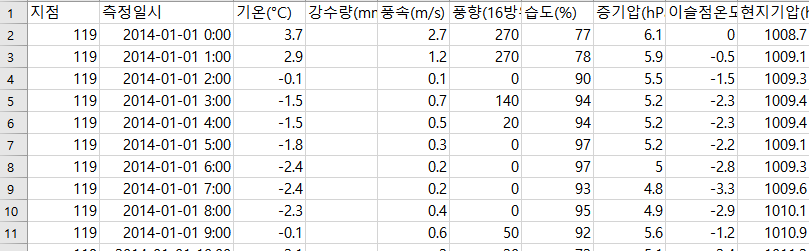
2. 데이터 전처리

본 프로젝트는 경기도권 2014년 ~ 2016년 대기오염물질과 종관관측기상데이터를 기반으로 분석을 실시한다. 두 데이터 모두 Open데이터이며 한국환경공단 AirKorea 및 기상청 기상자료 개방포털에서 얻을 수 있으며 CSV파일 형식이다. 데이터 전처리는 대기오염물질 데이터 전처리, 종관기상데이터 전처리, 대기오염물질과 종관기상데이터의 통합데이터 전처리로 진행한다. 대기오염물질데이터는 지역, 시간별로 분석 및 시각화 그리고 머신러닝의 기반자료로 사용되며 통합데이터는 지역, 시간별로 분석 및 시각화자료로 사용한다.

다음은 대기 오염물질 데이터와 종관 기상관측 데이터의 원 자료 형태이다.



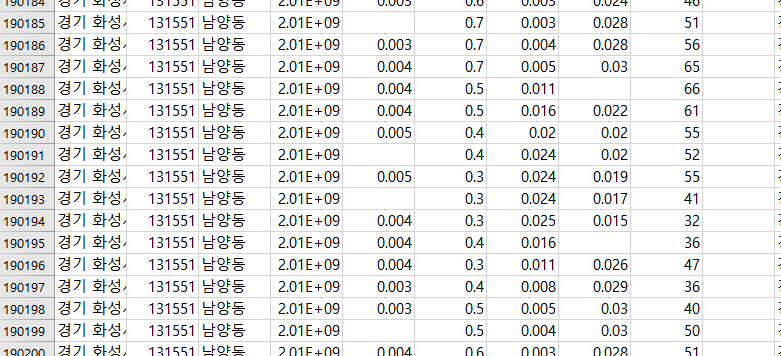
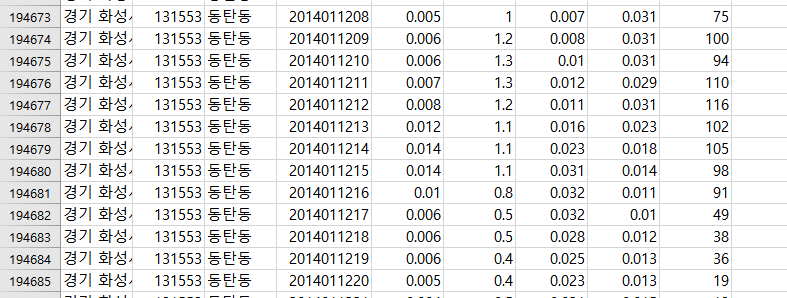
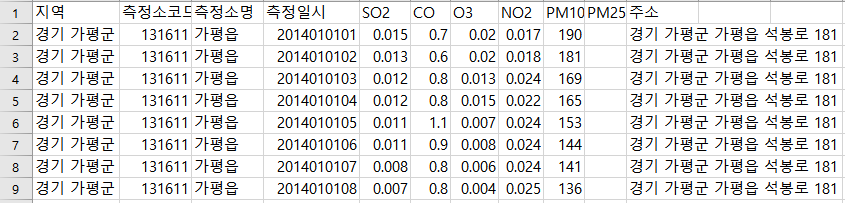
[그림1 – 대기오염물질 데이터.CSV(2014~2016) ]



[그림2 – 종관기상관측장비 기반 기상 데이터.CSV(2014~2016) ]

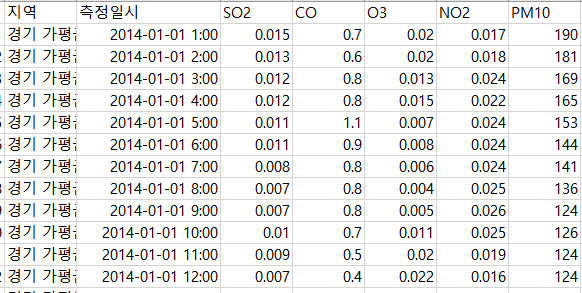
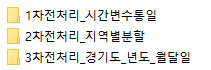
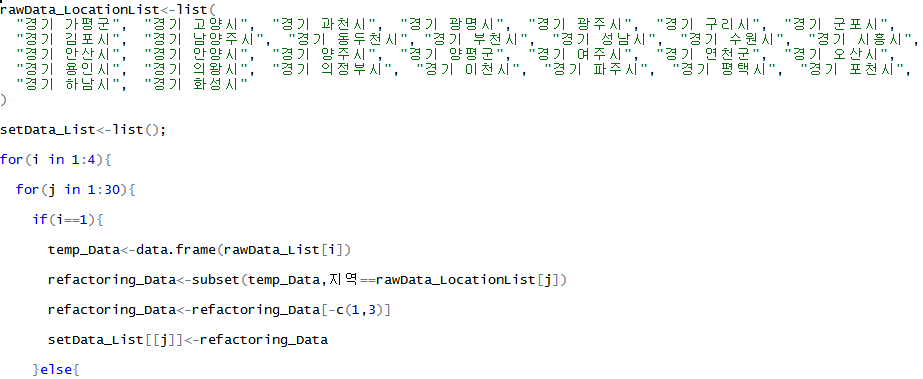
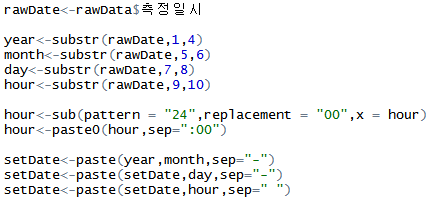
그림1의 대기오염물질 데이터는 다음 3가지의 문제점이 있었다.

1. 데이터의 결측치 존재
2. 측정일시 변수가 표준을 따르지 않음(YYYY-MM-DD-HH:MM)
3. 하나의 지역에 많은 측정소들이 존재함



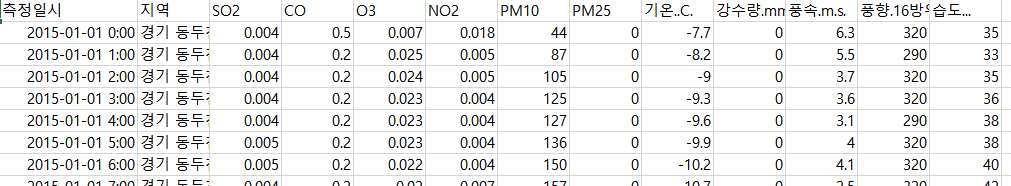
[그림3 – 대기오염물질 데이터의 전처리 요소 파악 ]

이 점을 해결하기 위해 측정일시 변수를 분해하여 표준형태를 따르는 Date변수를 생성하였다. 또한 결측치를 제거하고 하나의 시에 많은 측정소들의 측정값들을 시 별 평균값으로 대체하여 통합하였다.



[그림4 – 대기오염물질 데이터의 전처리]

마지막으로 정제된 대기오염물질 데이터와 종관관측기상 데이터를 R의 Merge함수를 이용하여 결합하였다.



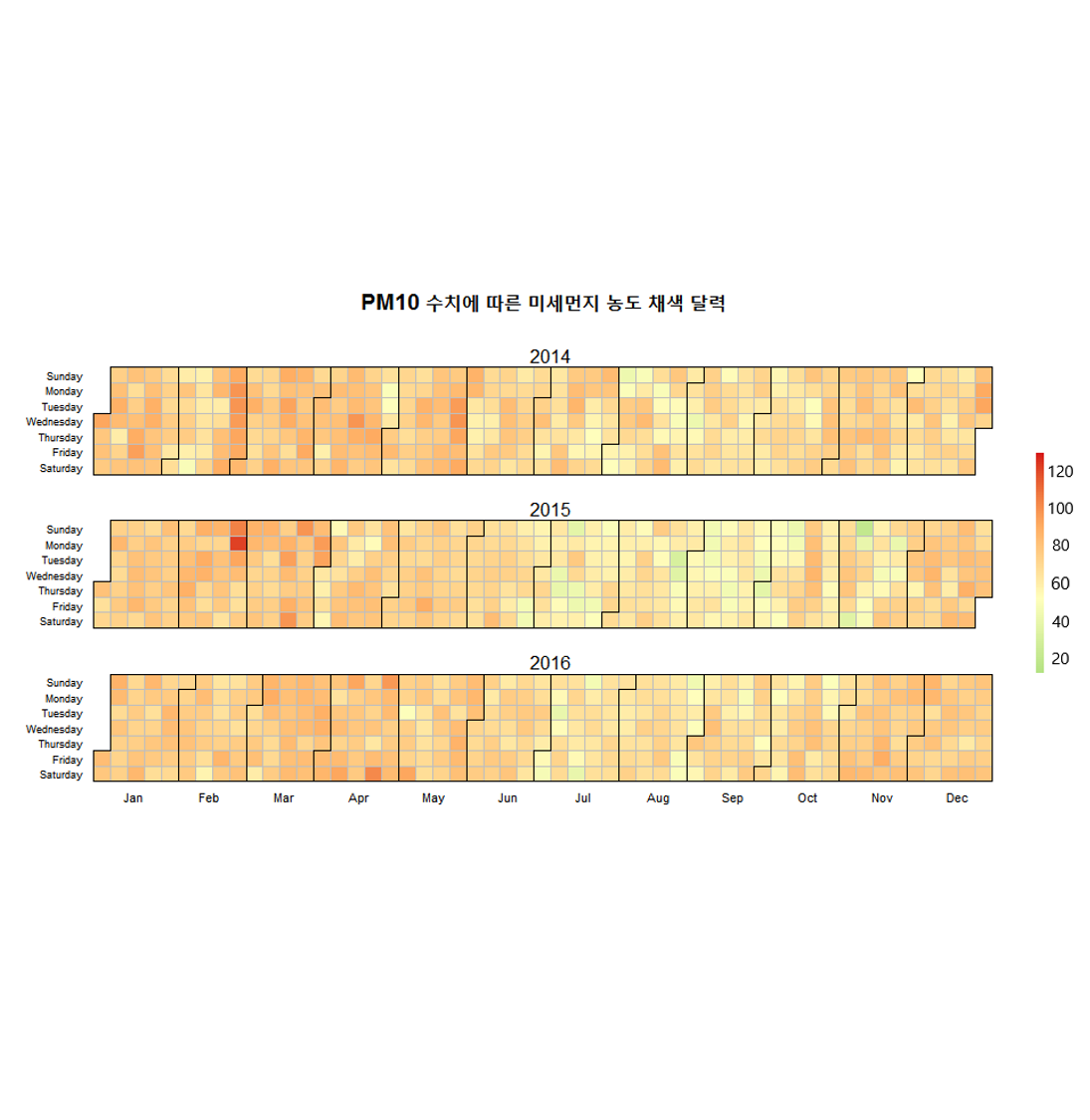
[그림5 – 통합 데이터의 전처리]

3. 데이터 분석

3.1 데이터 시각화

3.1.1 수치에 따른 HeatMap Calender

대기오염물질을 연도별로 나누어 분석해본 결과 미세먼저의 농도가 가장 두드러지게 수치가 높았다. 국내의 미세먼지 농도는 수치 60이상일 경우 ‘보통’의 경보가, 수치 90이상일 경우 ‘나쁨’의 경보가 발령된다. 다음은 2014년 ~ 2016년의 미세먼지 농도 수치를 월별로 나타낸 Grape이다.

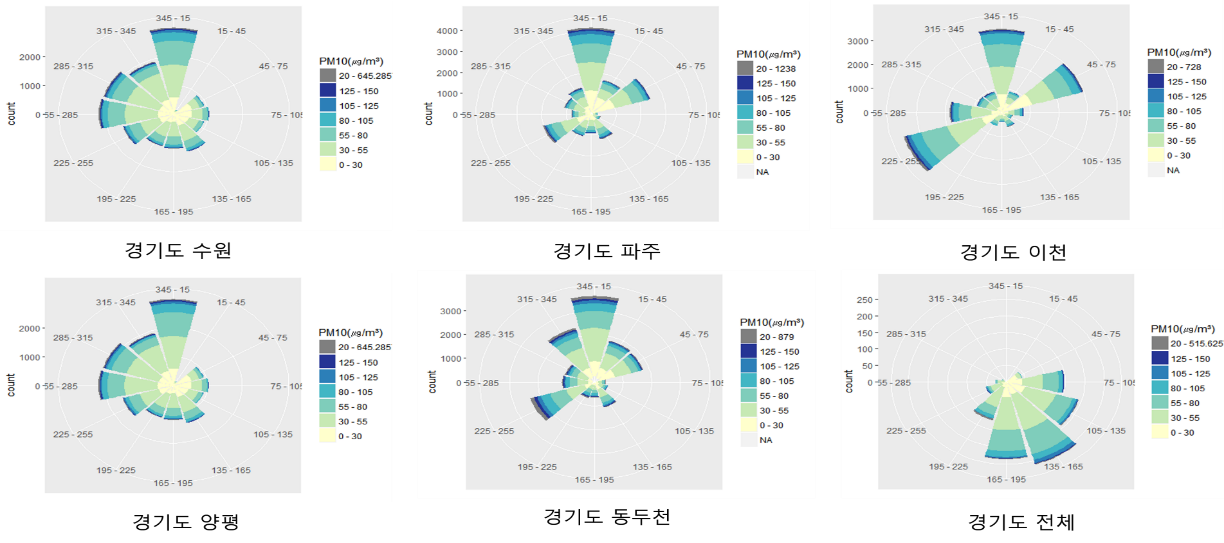


[그림6 – 미세먼지 농도 HeatMap Calender]

시각화한 결과 초록색의 ‘좋음’의 수치가 별로 없었다. 그만큼 최근 3년의 미세먼지 농도가 심각하다는 결과를 도출할 수 있었다. 또한 경기도의 미세먼지 농도는 겨울과 봄에 보다 더 높게 나오는 것으로 확인되었다.

* + 1. 풍향, 풍속, 수치에 따른 Wind Rose 시각화

대기오염물질 데이터와 종관기상데이터를 통합한 데이터를 토대로 Wind Rose시각화를 실시하였다. 다음 그림은 미세먼지의 농도를 풍향과 풍속에 따라 시각화한 결과이다.



[그림7 – 풍향과 풍속에 따른 미세먼지 농도 Wind Rose 시각화]

경기도의 미세먼지 농도를 동서남북의 4개의 지역의 각각 평균과 경기도 전체의 평균이다. 4개 지역의 각 평균은 대체적으로 북서풍이 불 때, 즉 겨울에 미세먼지 농도가 높아짐을 관측하였다. 하지만 경기도의 전체 지역의 평균은 남동풍이 불 때, 즉 봄에 미세먼지의 농도가 높아짐을 확인하였다.

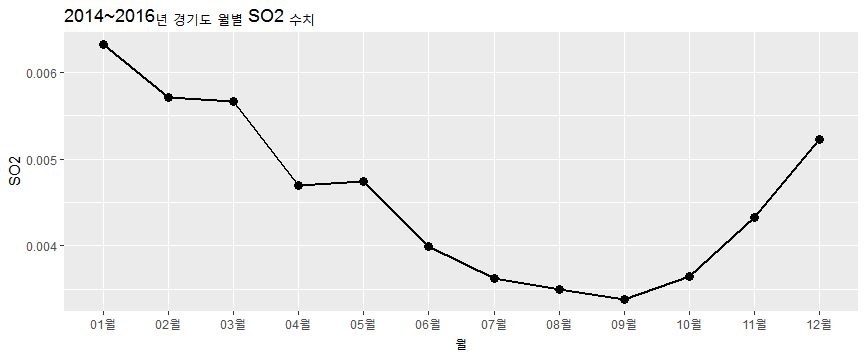
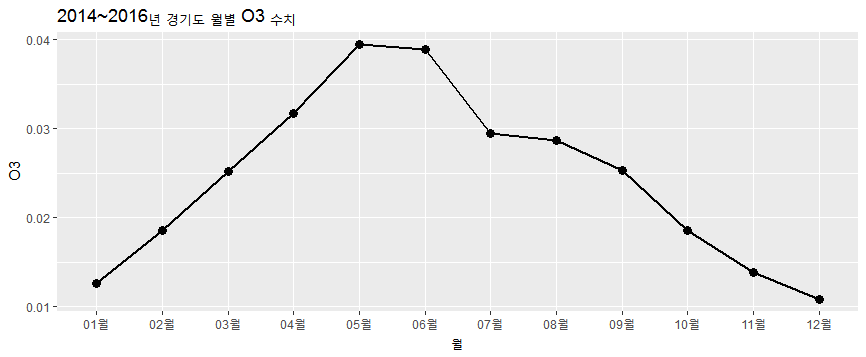
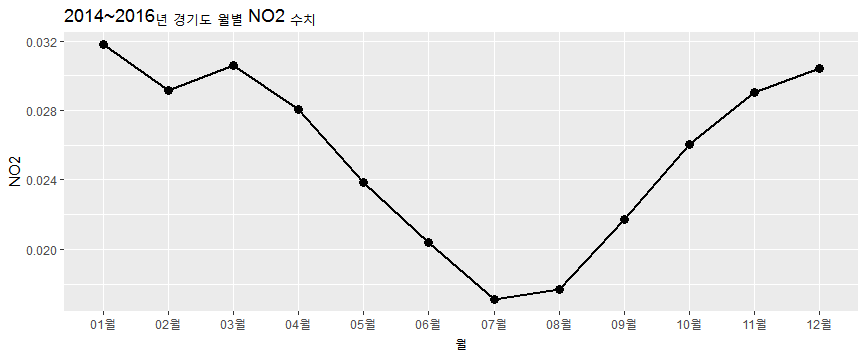
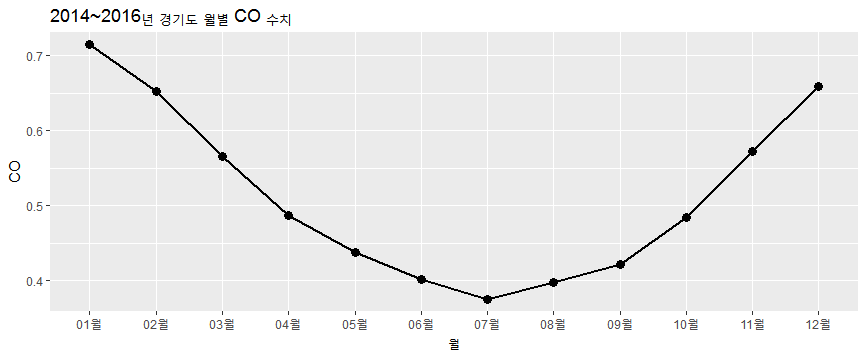
* + 1. 각 대기오염 물질에 따른 Chart 시각화

경기도의 3년간 대기오염물질을 요소별로 Chart Grape로 분석 및 시각화한 자료이다. 대기오염물질은 PM10(미세먼지), NO2, O3, CO, SO2 등 5개 요소이고 결과 분석은 다음 그림과 같다.



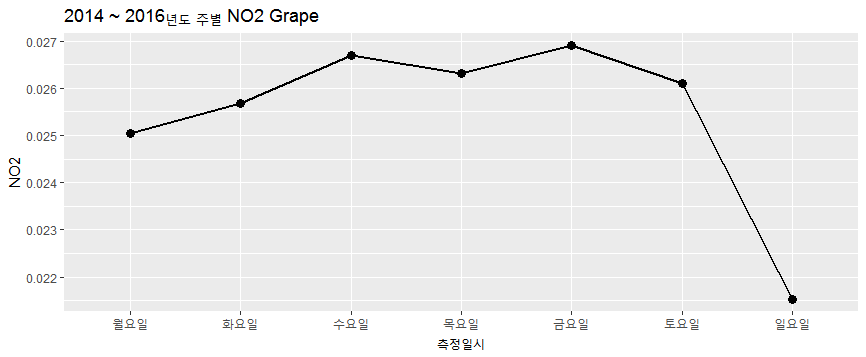
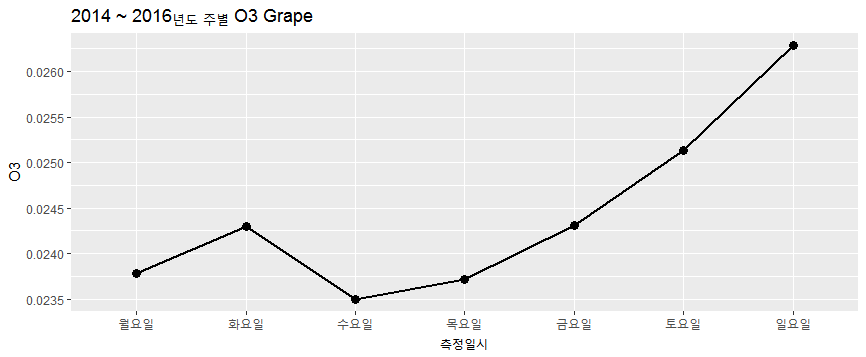
[그림8 – 경기도 미세먼지의 농도 시각화 그래프 ]

미세먼지의 경우 겨울이 다가옴에 따라 농도가 서서히 높아지다가 봄에 최고 수준의 농도를 보였다.



[그림9 – 경기도 대기오염물질 요소별 농도 시각화 그래프 ]

O3를 제외한 나머지 3개의 물질은 대체로 비슷한 패턴을 보였다. O3의 경우 NO2의 농도가 높아짐에 따라 감소하는 경향을 보인다.



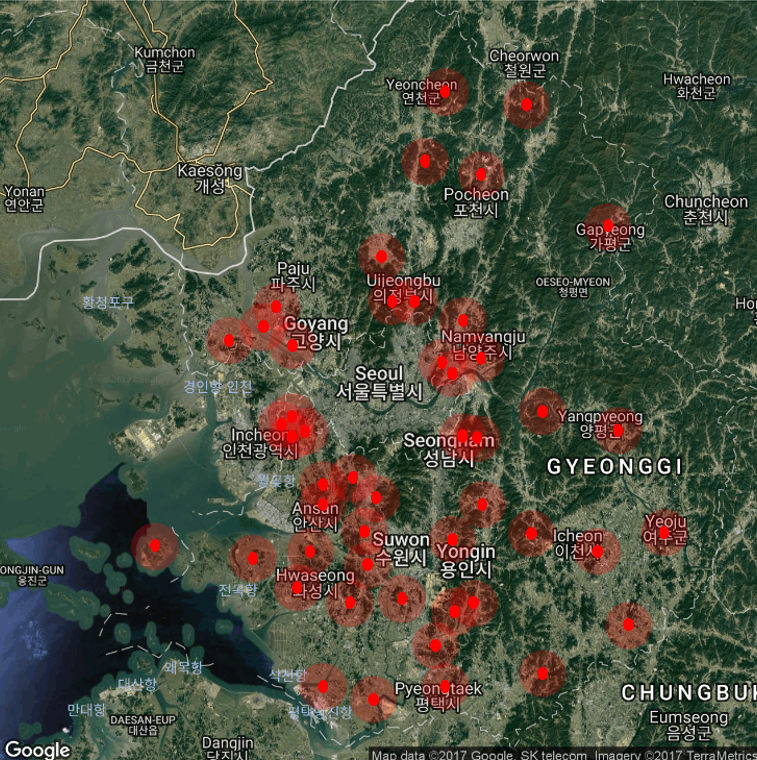
[그림10 – 경기도 주별 NO2, O3 농도 시각화 그래프 ]

NO2는 주로 자동차 주행 시 나오는 매연의 주성분이다. 위 분석 결과를 보면 차량 운행이 적은 주말에는 NO2의 농도가 낮고 O3의 농도가 높아짐을 확인할 수 있었다. 본 전처리 데이터는 대기오염물질 데이터를 지역별로 가지고 있음으로 필요시 분석하여 농도수치를 도출한다면 주간 차량운행 행정제도 마련에 유용할 것으로 보인다.

3.2 관측소 현황 공간 분석

경기도의 전체 대기오염물질의 관측소 위치 정보 및 관측범위를 구글지도 위에 표기하였다. 현재 우리나라는 환경부의 ‘인구 10만명 이상의 도시에 측정소를 설치해야 한다’ 는 지침을 따르고 있다. 지도위에 표시한 정보를 바탕으로 인구밀도와 결합하여 분석해본 결과 경기도는 현재 아래 나열된 지역에 추가적인 대기오염 관측소가 필요한 것으로 보인다.

|  |
| --- |
| **인구 밀도 10(인/ha) 이상 지역** |
| 남양주시 화도읍 – 103,221  광주시 오포읍 – 90,153  화성시 봉답읍 - 70,013  파주시 문산읍 – 47,118 |

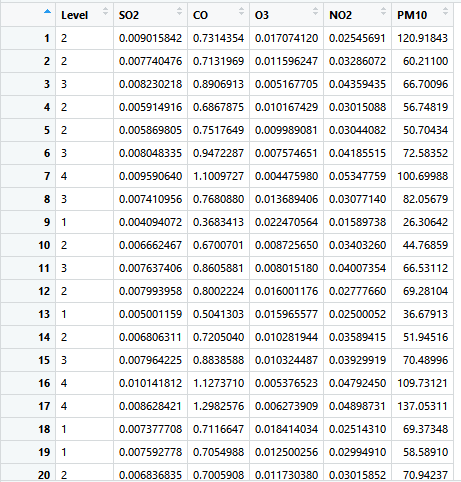
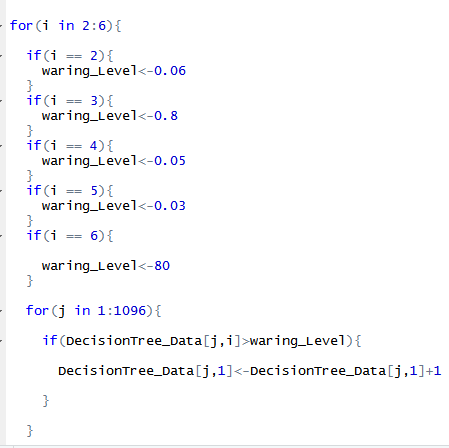


[그림11 – 경기도 관측소 현황 및 관측범위 시각화 ]

3.3 Machine Learning

3.3.1 Decision Tree에 따른 주요 대기오염물질 선정

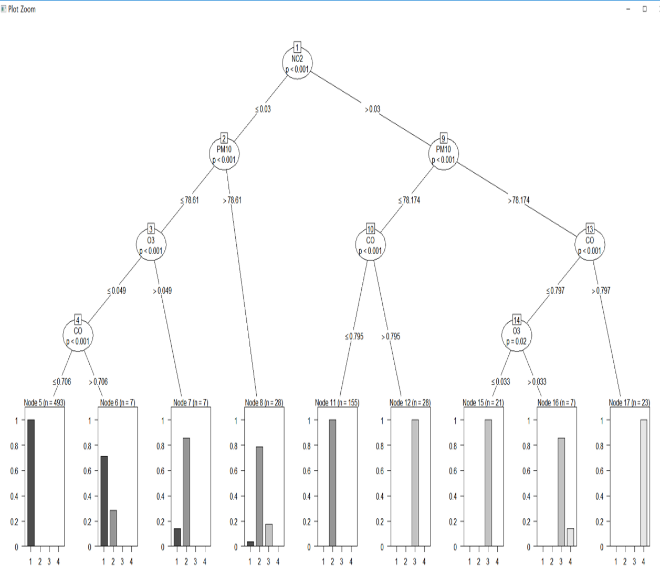
본 프로젝트는 Decision Tree를 수행하기 위하여 Level이라는 하나의 측정 변수를 만들었다. Level 변수는 5개의 각 대기오염물질이 국가의 ‘나쁨’ 수치의 이상일 경우 Counting 되며 이 변수를 종속변수로 정하여 Decision Tree 알고리즘을 실시하였다.

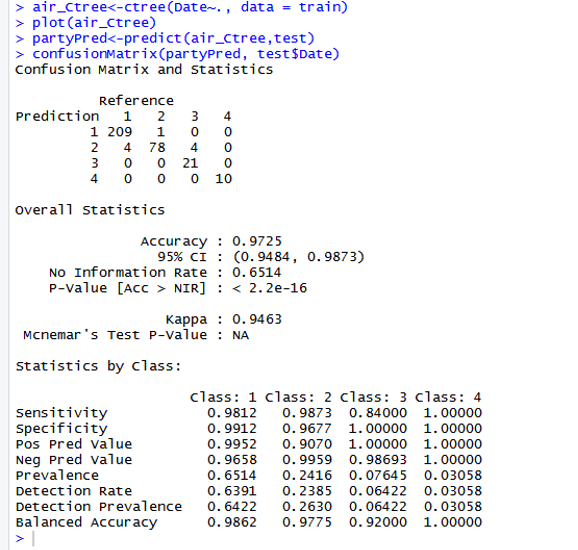


[그림11 – 경기도 관측소 현황 및 관측범위 시각화

[그림12 – Decision Tree 실행을 위한 종속변수 생성 ]

본 프로젝트는 Decision Tree알고리즘을 수행한 결과 NO2와 PM10이 유의한 변수로 검출되었다. 본 결과는 데이터를 7:3으로 나누어 재분석 및 검증결과 97% 신뢰도를 가졌다.

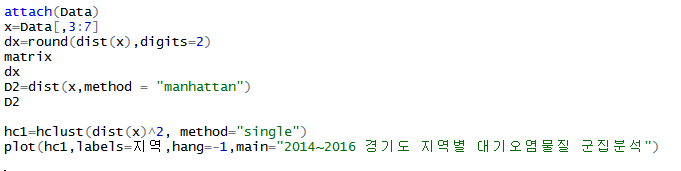
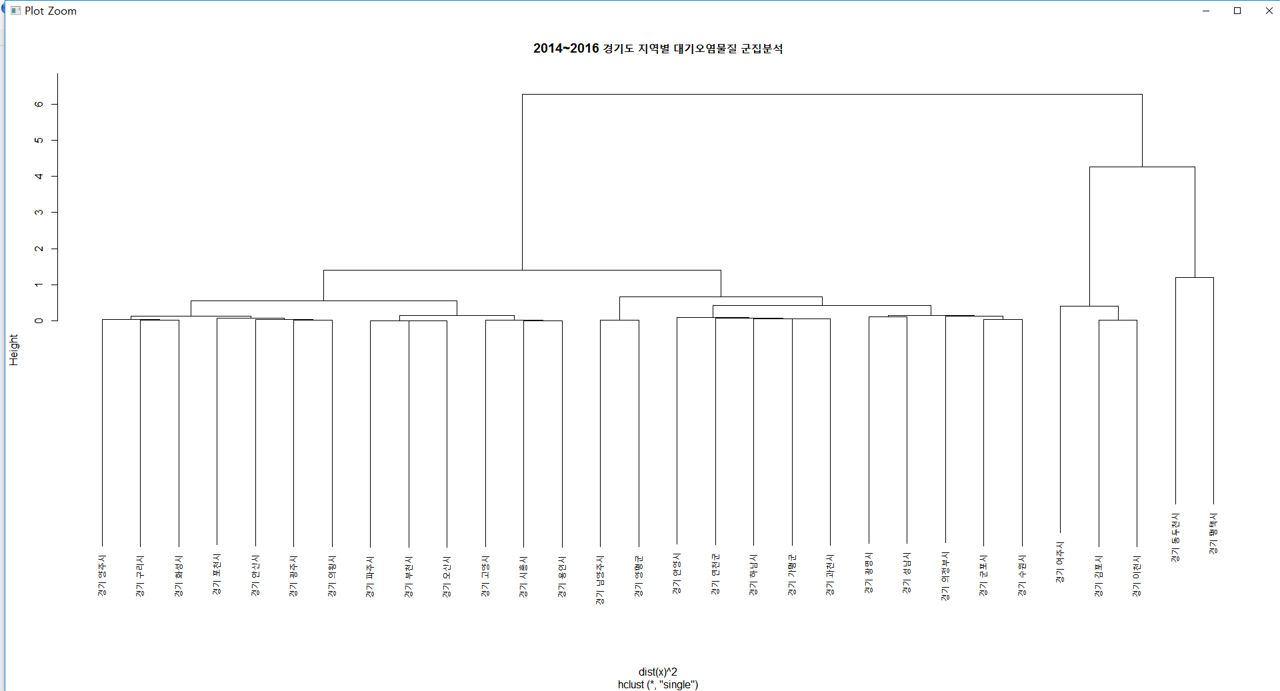




[그림13 – Decision Tree를 통한 유의한 변수 도출 및 신뢰도 ]

3.3.2 경기도 지역 군집 분석

경기도는 총 31개의 도시로 이루어져 있다. 본 프로젝트는 유사한 대기오염물질 속성을 가진 집단의 개수를 파악하기 위하여 31개의 도시의 대기오염물질 요소별 평균값을 기반으로 군집분석을 실시하였고 분석결과 총 3개의 군집이 확인되었다.



1

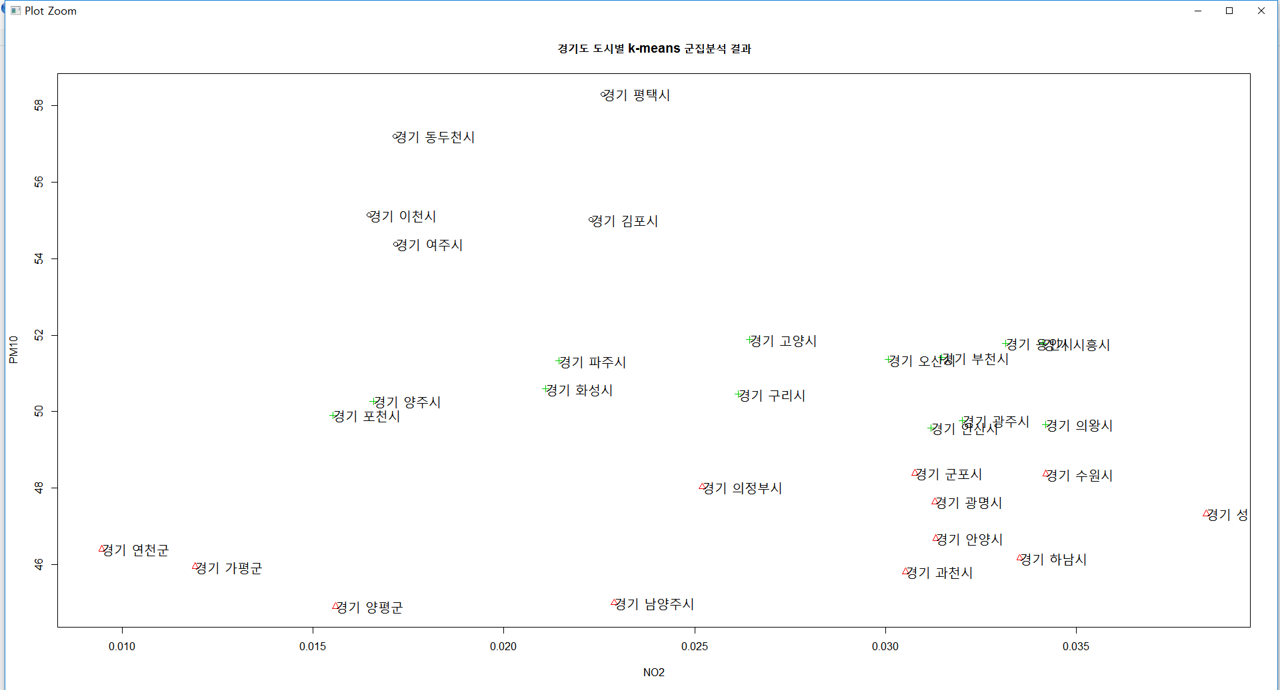
2

3

[그림14 – 경기도 31개의 도시 군집분석 결과 ]

3.3.3 k-means 군집분석

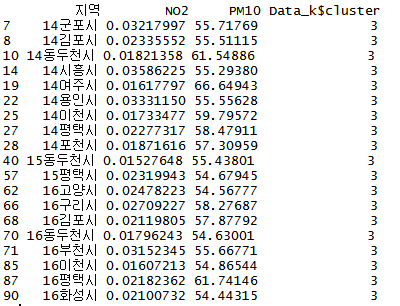
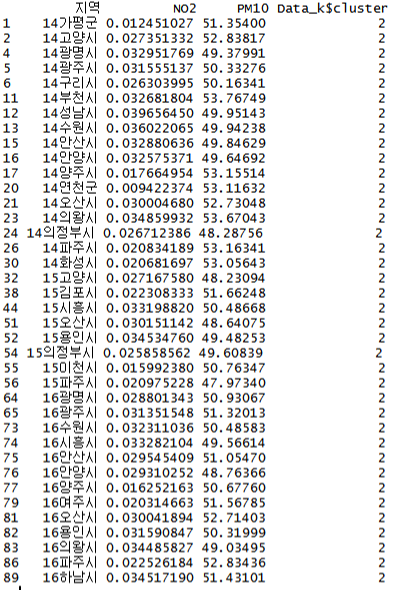
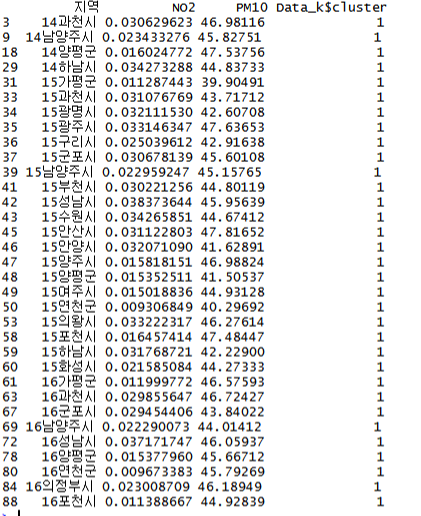
본프로젝트는 3.3.1과 3.3.2의 분석결과에 따라 경기도의 주요 대기오염물질은 NO2와 PM10이 유의한 변수임과 31개의 도시는 3개의 군집으로 비슷한 성질을 가지고 있음을 확인하였다. 이를 바탕으로 X, Y축에 유의한 변수 NO2와 PM10을 대입 후 3개의 군집으로 k-means 분석을 실시하였다.



[그림14 – 경기도 지역 k-means 분석 ]

분석 결과 동두천, 평택, 이천, 여주, 김포는 PM10이 가장 높은 군집을 이루는 것을 확인할 수 있었다. 반대로 수원과 하남 등의 도시는 NO2가 높은 군집을 이루는 것을 확인 할 수 있었다. 위 결과는 경기도의 도시 별 2014년 ~ 2017년 평균데이터의 k-means 분석결과이다.

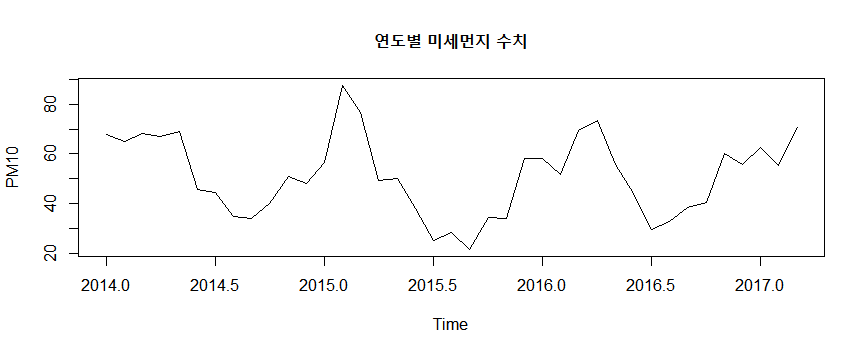
다음은 2014년, 2015년, 2016년 연도별로 도시의 평균을 요소로 두어 총 93개의 도시 별 평균데이터를 k-means 분석한 결과이다. 3개의 군집요소의 분석결과에 따라 각 도시의 연도별 주요 대기오염물질의 변화추이를 관측할 수 있었다.



[그림14 – 연도별 경기도 지역 k-means 분석 ]

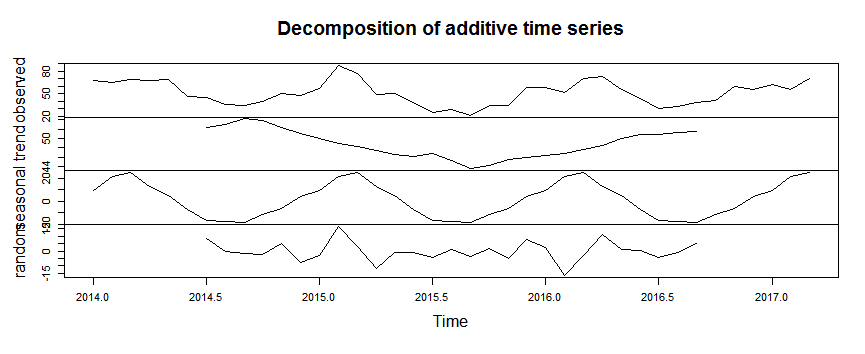
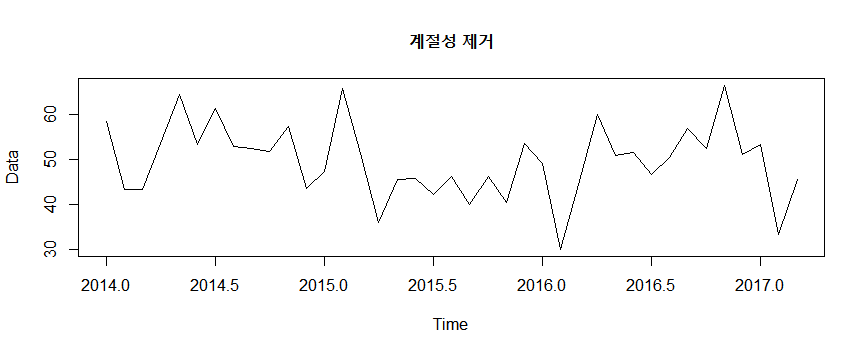
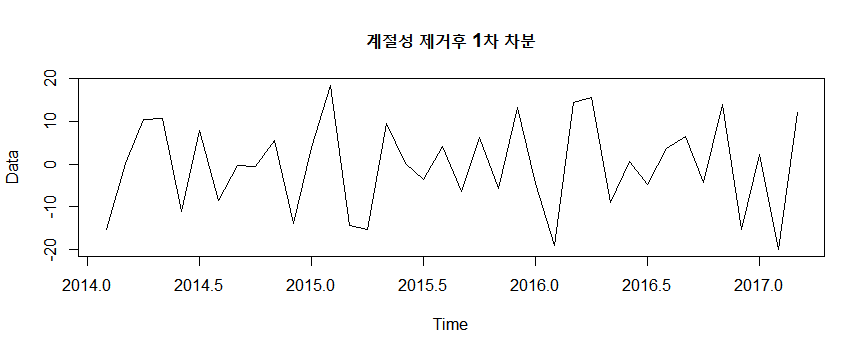
3.3.4 ARIMA 모형 시계열 데이터 분석 및 예측

본 프로젝트는 2014~2016년 미세먼지 데이터를 기반으로 ARIMA 시계열 데이터 분석을 통하여 향후 미래먼지의 수치를 예측 및 시각화 하였다. 다음은 경기도의 3년간 미세먼지의 수치를 Grape화 한 것이다.



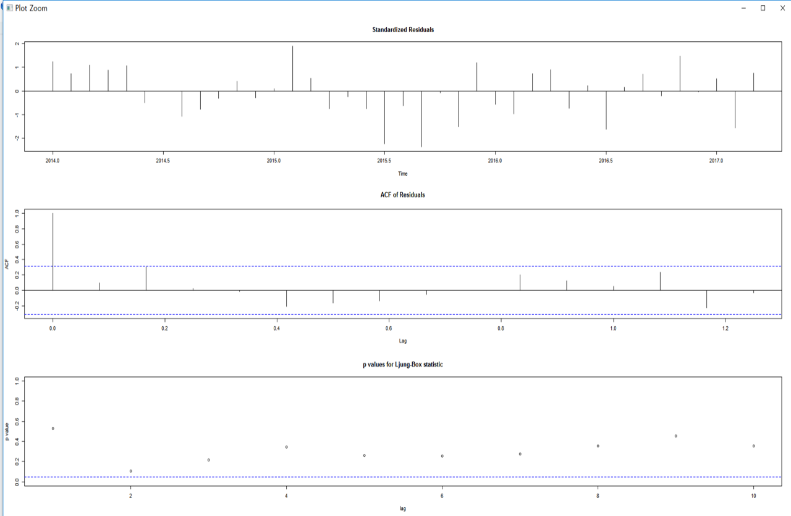
[그림15 – 연도별 경기도 미세먼지 Grape ]

ARIMA 시계열 분석을 위해서는 데이터가 정상시계열이야 함으로 본 기반 데이터의 요소를 분해 및 시각화 해보았다. 그 결과 계절성과 경향성을 보였고 정상시계열화를 위하여 기반 데이터에서 계절요소를 제거한 후 1차 차분결과 정상시계열 데이터 형태를 이루었다.

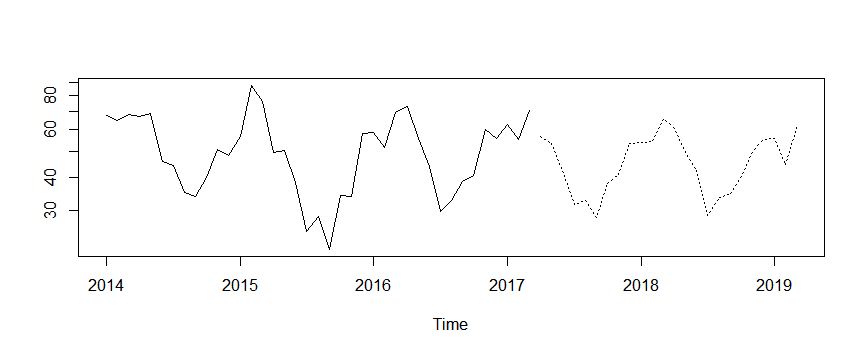
[그림16 – ARIMA 모형 적용을 위한 데이터 정상시계열화]

정상 시계열화 한 데이터를 기반으로 여러가지 ARIMA모형을 도출하였고 그 결과 ARIMAs (0,0,1)(1,0,2)12 의 모형이 가장 적합하였다. 다음은 도출된 ARIMA모형의 단위검증 결과이다.



[그림17 – 도출된 ARIMA 모형의 검증 결과]

검증결과 본 모델은 잔차의 뚜렷한 패턴 과 자기상관성이 존재하지 않음 및 White Noise 형태를 확인할 수 있었다. 다음은 도출된 모델을 토대로 미래의 미세먼지 추이를 예측한 결과이다.



4. 결론(기대효과)

본 프로젝트는 다음과 같은 3가지의 기대효과가 있다. 첫째로, 인구밀도와 대기오염물질의 지역특성, 공간분포상의 취약지를 고려하여 추가 대기환경 측정소 설치지역을 파악하였다. 이 데이터는 정부의 미세먼지와 같은 대기오염의 예보, 경보제 운영인프라 확충에 쓰일 수 있으며 추가 측정망 설치에 활용할 수 있다. 둘째로, 지역 별, 세부 시간 별 대기오염물질 변화 추이를 분석하여 유의미한 결과를 도출하였다. 또한 유의한 오염물질 변수를 파악하여 비슷한 특징을 가지는 지역을 군집으로 연도별 변화추이를 파악하였다. 이 데이터는 정부가 대기오염 관련 대책을 수립할 때 지역 군집을 바탕으로 효율적인 연계 대책을 수립할 수 있을 것이다. 마지막으로, 다양한 머신 러닝 방법을 이용하여 가장 성능이 좋은 모델을 활용 및 미래 시점 미세먼지의 추이를 예측하였다. 이 데이터는 한국형 미세먼지 예보모델의 개발 및 예보 권역 별 상세모델 구축의 자료로 활용될 수 있다.

6. 참고자료

예측방법론(저자 : 이긍희 외 2인 / 출판사 : 한국방송통신대학교 출판문화원)

시계열 분석과 예측(저자 : 이우리 / 출판원 탐진)

R을 이용한 데이터 처리&분석 실무(저자 : 서민구 / 출판사 : 한빛미디어)

R을 이용한 데이터 시각화(저자 : 전희원 / 출판사 : 자유아카데미)

R을 이용한 데이터 마이닝(저자 : 박창이 외 4인 / 출판사 : 교우사)

SAS 다변량 자료분석 입문(저자 : 강현철 외 3인 / 출판사 : 자유아카데미)

R좀 R려줘(저자 : 김승욱 / 출판사 : 느린생각)