빅데이터 활용

미세먼지와 호흡기 질환 연관성 분석

**2020. 4.**

▮ 요 약

최근 국내 미세•초미세먼지를 포함한 대기오염 물질에 따른 피해가 급증함에 따라, 해당 오염물질이 국민 건강에 미치는 영향에 대해 사회적인 관심이 증대되고 있다. 본 연구에서는 미세먼지와 호흡기 질환 간의 연관성을 분석하자는 의도에서 출발하여, 미세먼지 이외의 대기오염 물질 및 기상 요인이 호흡기 질환에 영향이 있는지도 분석하였다.

본 연구를 위하여 기상데이터 포털과 국민건강보험공단으로부터 2016년부터 2018년까지 3년간의 공공 개방 데이터를 수집하였다. 설명 변수로는 2016년 ~ 2018년의 대기오염 물질(PM10, PM2.5, O3, NO2, CO, SO2)의 측정값과 기상요인(기온, 풍속, 풍향, 습도)를, 반응변수로는 동기간 국내 호흡기 질환 진료내역 및 지역별 인구자료를 통해 계산한 호흡기질환 발병률을 사용하여 분석하였다.

기상요인 및 대기오염물질과 호흡기질환 발병률 간의 시계열 패턴분석을 통해 변수간 연관성을 추정해보았다. 각 설명변수와 반응변수 시계열 자료를 STL 방법으로 분해하여 패턴을 분석하고, 유사점과 차이점 비교를 통해 변수 간의 연관성에 대해 추정하였다. 22

또한, 호흡기 질환 발병율과 높은 상관성을 보이는 기상•대기 요인을 설명 변수로 하여 회귀분석을 진행하였다. 본 회귀분석을 수행하기위해 변수 간 요일요인 및 지연효과를 고려하였다. 분석 결과, 설명 변수 중 O3가 호흡기질환 발병률에 미치는 영향이 통계적으로 가장 유의미하다는 것을 확인하였으며 그 다음으로는 PM10, 평균 현지기압 순으로 나타났다.

분석 결과, 호흡기 질환 발병률에 가장 영향력이 높은 요인은 O3와 PM10으로 나타났고, 한편 크게 영향을 미치지 않을 것으로 보였던 평균 현지기압도 통계적으로 호흡기질환 발병률에 영향을 미치는 것으로 나타났다. 이에 대한 추가적인 연구가 필요할 것으로 보인다.

주제어 : 미세먼지, 호흡기질환, 대기오염물질, 기계학습, 빅데이터

▮차 레

[제1장 연구 개요 1](#_Toc38986583)

[1. 배경 및 필요성 1](#_Toc38986584)

[2. 내용 및 범위 2](#_Toc38986585)

[3. 수행방법 및 절차 3](#_Toc38986586)

[제2장 데이터 수집 및 정제 4](#_Toc38986587)

[1. 데이터 수집 4](#_Toc38986588)

[2. 데이터 전처리 5](#_Toc38986589)

[제3장 대기 오염 물질 분석 8](#_Toc38986590)

[1. 미세먼지 8](#_Toc38986591)

[2. 초미세먼지 11](#_Toc38986592)

[3. SO2 13](#_Toc38986593)

[4. NO2 17](#_Toc38986594)

[5. O3 20](#_Toc38986595)

[6. CO 23](#_Toc38986596)

[제4장 기상 분석 26](#_Toc38986597)

[1. 기온 26](#_Toc38986598)

[2. 기압(평균현지기압) 29](#_Toc38986599)

[3. 풍속 31](#_Toc38986600)

[4 풍향 34](#_Toc38986601)

[5 강수 36](#_Toc38986602)

[6 강설(최심신적설) 39](#_Toc38986603)

[제5장 진료내역 분석 42](#_Toc38986604)

[1. 발병률 분석 42](#_Toc38986605)

[2. 기상요인, 대기오염물질에 따른 발병률 분석 45](#_Toc38986606)

[제6장 군집분석 51](#_Toc38986607)

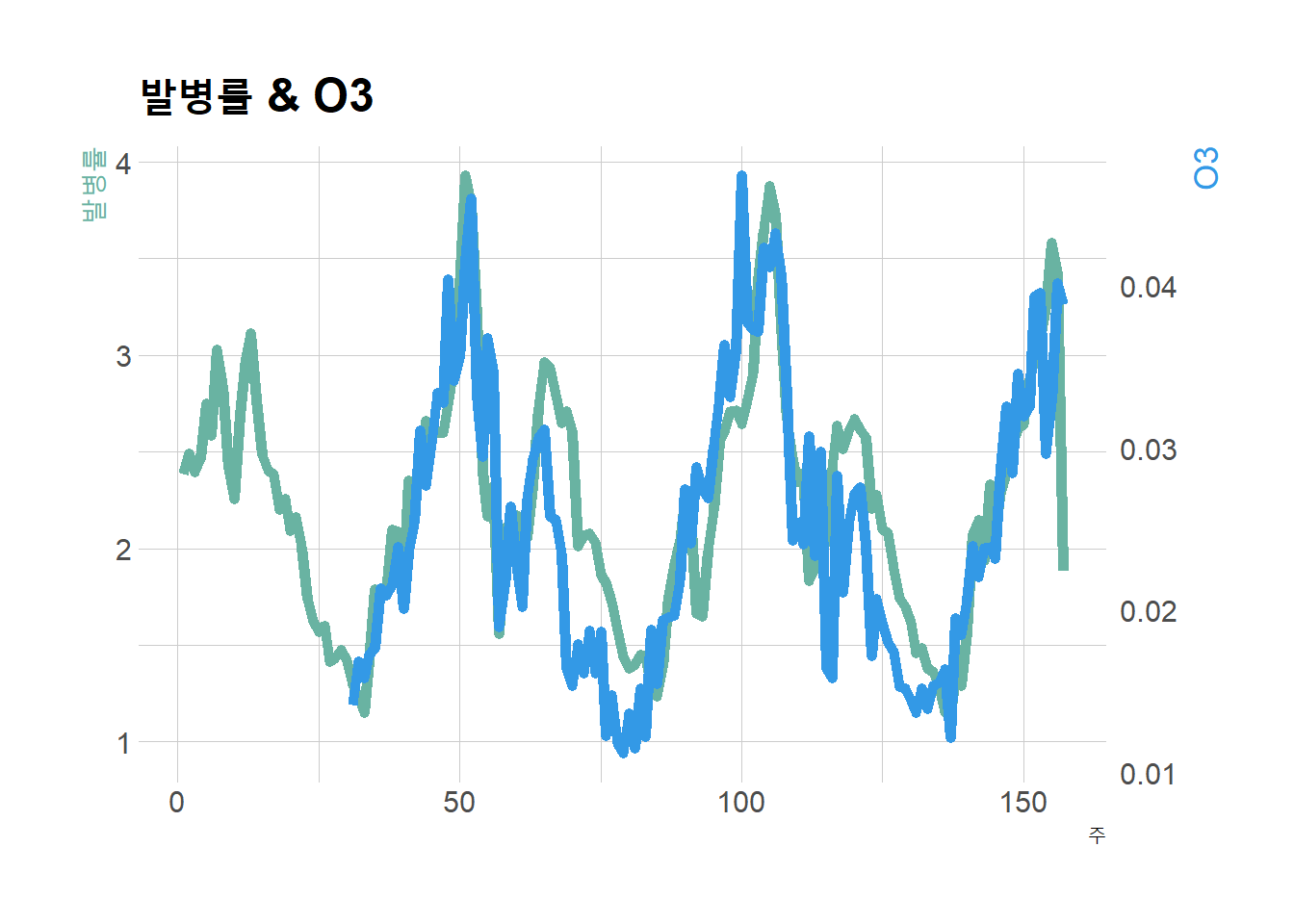
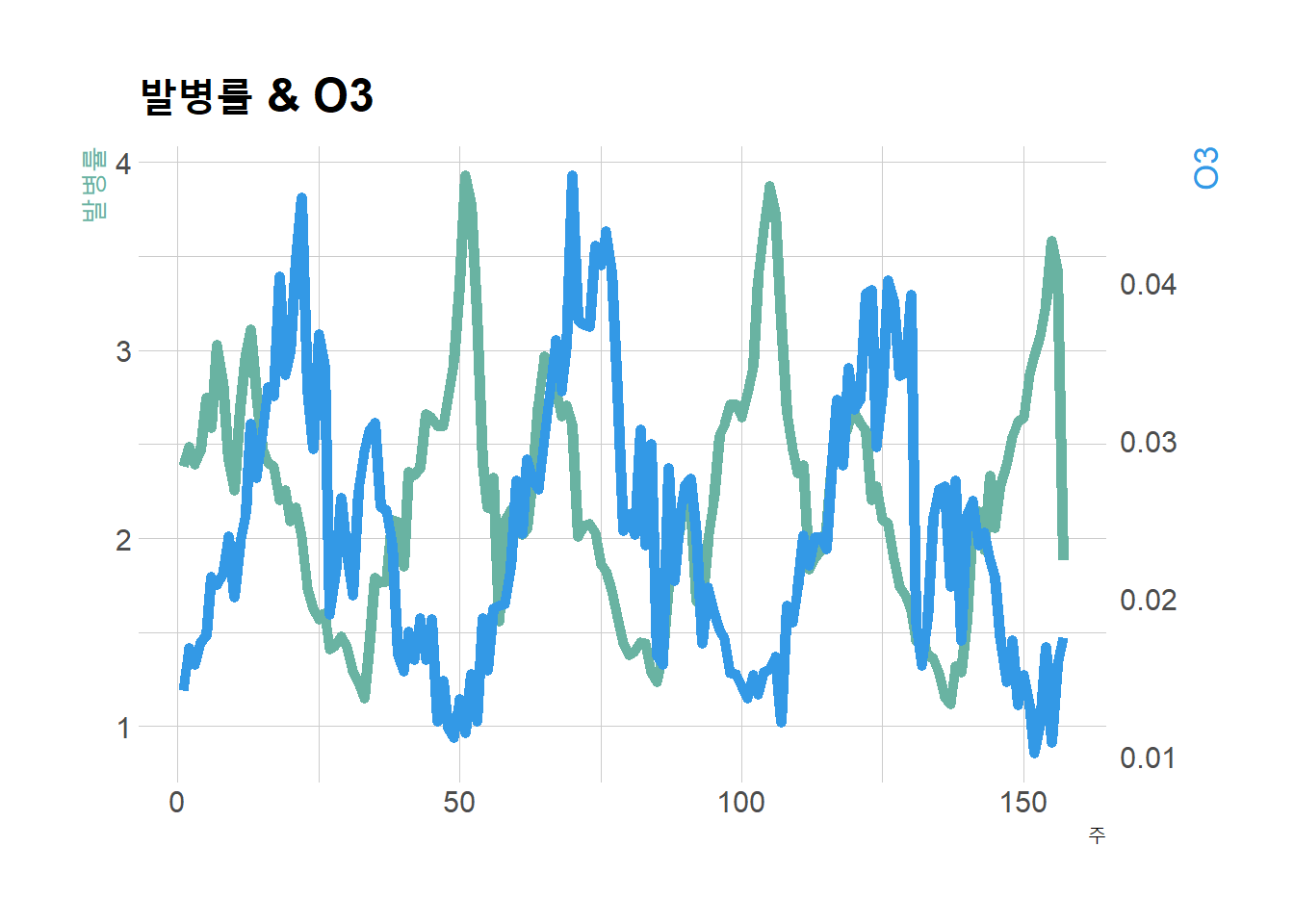
[1. 지역별 요인 비교 51](#_Toc38986608)

[2. 군집화 52](#_Toc38986609)

[제7장 회귀분석 54](#_Toc38986610)

[1. 요인과 발병률 간 상관관계 확인 54](#_Toc38986611)

[2. 지연효과 적용 56](#_Toc38986612)

[ 56](#_Toc38986613)

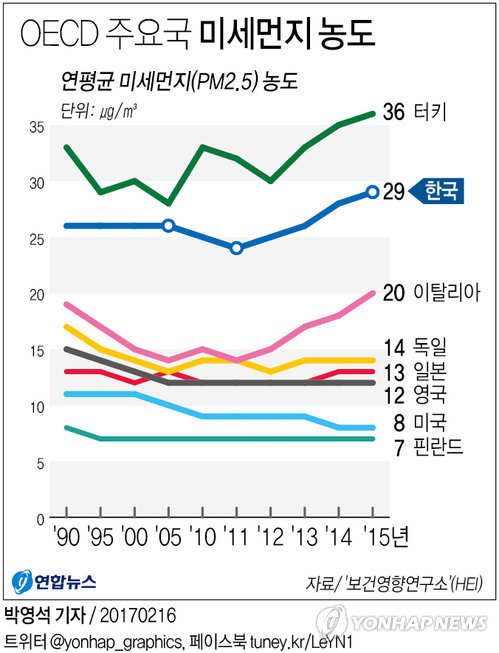
[제8장 결론 및 제언 58](#_Toc38986614)

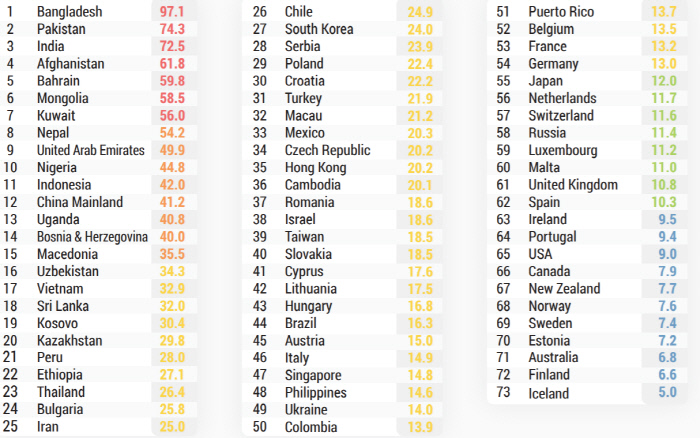
[1. 결론 58](#_Toc38986615)

[2. 제언 58](#_Toc38986616)

# 연구 개요

## 배경 및 필요성

최근 미세먼지로 인한 사회적 피해가 날로 증가하고 있으며 세계적인 문제로 대두되고 있다. 세계 보건기구(WHO)에서는 미세먼지를 1군 발암 물질로 지정하여 그 심각성과 위험성을 경고하고 있다. 특히 한국은 경제협력개발기구(OECD) 회원국 중 초 미세먼지 농도가 높은 나라 2위에 올랐으며, 서울은 세계 수도 62곳 중 공기의 질이 27번쨀 나쁜 도시로 꼽혔다. 이에 정부는 2016년부터 9대 국가전략 프로젝트 중 하나로써 미세먼지 R&D사업을 지정하고 가장 많은 예산을 투입하고 있다.



[ 그림 1 ] OECD 주요국 및 세계 수도 미세먼지 농도

또한 호흡기 원인에 대한 뉴스를 크롤링 하여 단어 분류를 한 결과 미세 먼지가 가장 높은 빈도수로 추출이 되어 호흡기 질환의 원인이 미세먼지라는 것이 대중적인 인식임을 확인할 수 있다. 이에 따라 미세 먼지가 호흡기 질환에 실제로 영향을 미치는지 확인해볼 필요성이 제고된다.

[ 그림 2 ]호흡기 질환 원인관련 뉴스의 워드 클라우드

## 내용 및 범위

요구사항명세서에 따른 내용은 다음과 같다.

* 1. 호흡기 질환에 영향을 끼치는 기상요인 전처리
     + 미세먼지와 호흡기 질환관련 선행연구 분석 및 데이터 특성파악
     + 데이터수집
       - 국내 미세먼지 관련 대기 측정값 등 관련 데이터수집

\* 측정소정보, 측정값 등

* + - * 기상청, 통계청에서 제공하고 있는 관련 공공데이터
      * 국민건강보험공단에서 제공하는 진료내역 데이터
    - 데이터 전처리
      * 주소 및 위치정보 정제, 데이터 병합 및 분할 등 데이터 클린징
      * 진료내역 중 호흡기 질환을 진단받은 데이터 분리
    - 데이터 융합을 통한 분석 데이터셋 구축
      * 각 지역 인구데이터를 활용하여 진료건수 보정
      * 데이터병합, 조인을 통한 통합 분석 데이터셋 구축
  1. 통계분석을 통한 미세먼지와 호흡기 질환 간연관성분석
     + 지역별시계열 분석수행
       - 측정값, 기상데이터 중심으로 시계열 분석으로 패턴도출
       - 지역별 특정 패턴 및 특징 파악
       - 요인별 패턴을 비교하여 상관성 파악
     + 통계분석 분석 결과 도출
       - 분산분석을 통한 지역별 기상요인 차이 분석
       - 상관분석, 회귀분석 등을 통한 영향력 분석

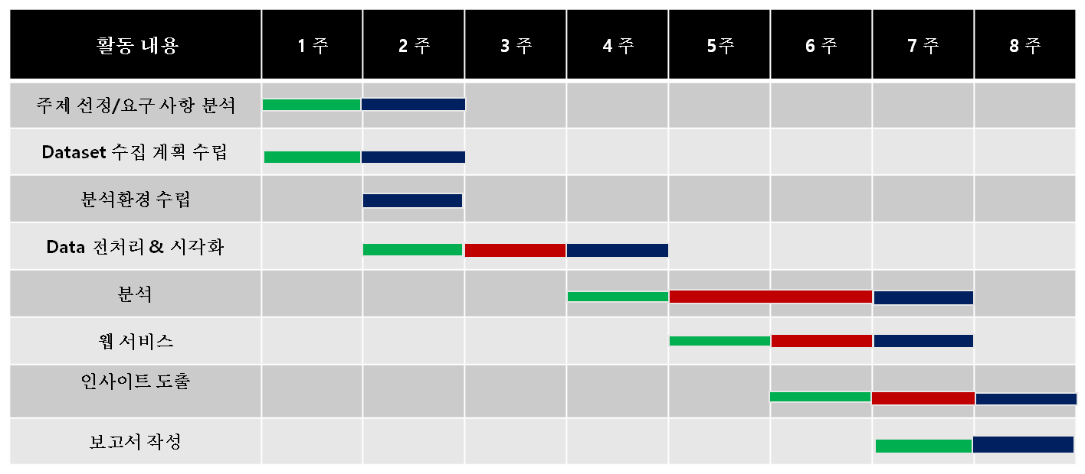
1. 도출된 요인별 호흡기 질환 연관성 파악 및 대책 수립
   * 다양한 통계 분석을 통해 도출된 결과를 바탕으로 인사이트도출
     + 통계분석, 상관분석, 회귀분석 등 분석결과를 바탕으로 정책 수립 활용 가능한 인사이트제공
   * 요인별 계량화 된 영향력을 통한 방안제시
     + 호흡기 질환 발생에 영향력이 높은 요인들의 우선순위를 바탕으로 구체적 저감 방안작성

또한, 연구의 시간적 범위는 2016년 ~ 2018년이며, 실제 이 기간을 중심으로 데이터 특성을 감안하여 융복합 하였다. 이 과정에서 데이터의 공통적 기간을 고려 하여 융합한 결과 2016년 ~ 2018년의 데이터셋이 만들어져 분석하였으며 범위는 전국이었다.

## 수행방법 및 절차

호흡기 질환과 미세 먼지 간 영향을 확인하기 위해 기상청과 국민건강 보험공단에서 다양한 관련 자료를 수집하고 빅데이터 융합을 통해 데이터셋을 구축하여 각 요인들을 분석한다.

데이터 전처리를 위해 시간별, 지역별로 분리되어 있는 데이터의 데이터 타입을 문자형, 숫자형으로 통일하고 결측치와 이상치 보정한다. 그 후 시도 코드를 부여하여 그룹바이(groupby)로 묶는다.

정제된 데이터를 시각화하여 발병률과 각 요인들의 추세를 비교하거나, 유사한 추세를 보이는 지역별로 군집 분석을 한다. 그 후 각 요인들이 상관이 있다고 판단되면 회귀 분석을 실행한다.

[ 그림 3 ] 연구 일정

# 데이터 수집 및 정제

## 데이터 수집

10종의 데이터를 각 기관으로부터 수집하였다. 시계열적 데이터와 공간적 데이터를 각각 수집하였으며 수집된 데이터를 융합하여 데이터셋을 구축하였다.

|  |  |  |
| --- | --- | --- |
| **데이터 명** | **주 요 속 성** | **출 처** |
| 대기오염 측정망 정보 | * 전국 321개 측정소의 기상 및 대기측정 자료 * 관측일시, PM10, O3, NO2, CO, SO2등 * 기온, 풍향, 풍속, 습도 | 한국환경공단, 에어코리아 |
| 기상 정보 | * 전국 방재기상관측(AWS), 종관기상관측장비(ASOS) * 관측일시, 기온, 풍속, 풍향, 습도 등 기상 정보 | 기상청 기상자료개방포털 |
| 진료내역 정보 | * 국민건강보험 가입자 중 요양기관(병/의원 등)으로 부터의 각 연도별 수진자 100만 명에 대한 정보 * 기본정보(성, 연령대, 시도코드 등)와 진료내역(진료과목코드, 주상병코드, 요양일수 등) | 국민건강보험공단 |

[ 표 1 ] 데이터 속성 및 출처

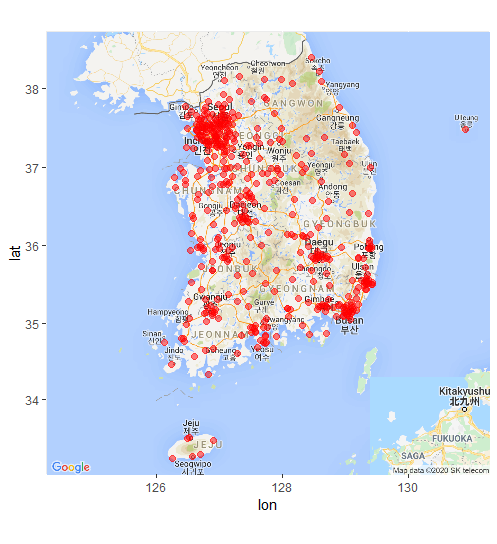
## 데이터 전처리

데이터 전처리는 시간별, 지역별 분리되어 있는 데이터를 문자형, 수치형 등 자료형을 일치시키고 결측치, 이상치(Outlier)의 제거, 보정 등 데이터를 정제하는 과정이다. 그리고 측정소별, 시기별로 나누어진 데이터셋을 시점 별 설명 변수와 지역별 설명변수로 설정하기 위해 시점 데이터와 지역 데이터를 조인(Join)하였다.

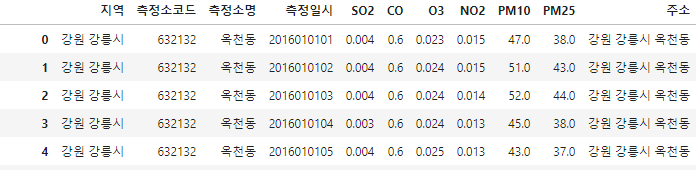
2.1 대기오염 물질 데이터 전처리

먼저, 전국의 각 대기오염 측정소의 정보를 한국환경공단에서 실시간 제공하는 대기오염도 정보(http://www.airkorea.or.kr)를 웹에서 수집하였다. 제공하는 정보는 전국의 측정망의 측정소 정보와 측정소에서 확정된 PM10, PM2.5, O3, NO2, CO, SO2의 관측값이다.

취득한 442곳 대기오염 측정소의 주소를 정제했다. 또한 2016년 ~ 2018년간 전국의 미세먼지 확정자료를 수집하고 일별로 통합하여 8개 필드, 18.631개 레코드의 데이터셋을 구축하였다.



[ 그림 4 ] 전국 대기오염 측정소 현황

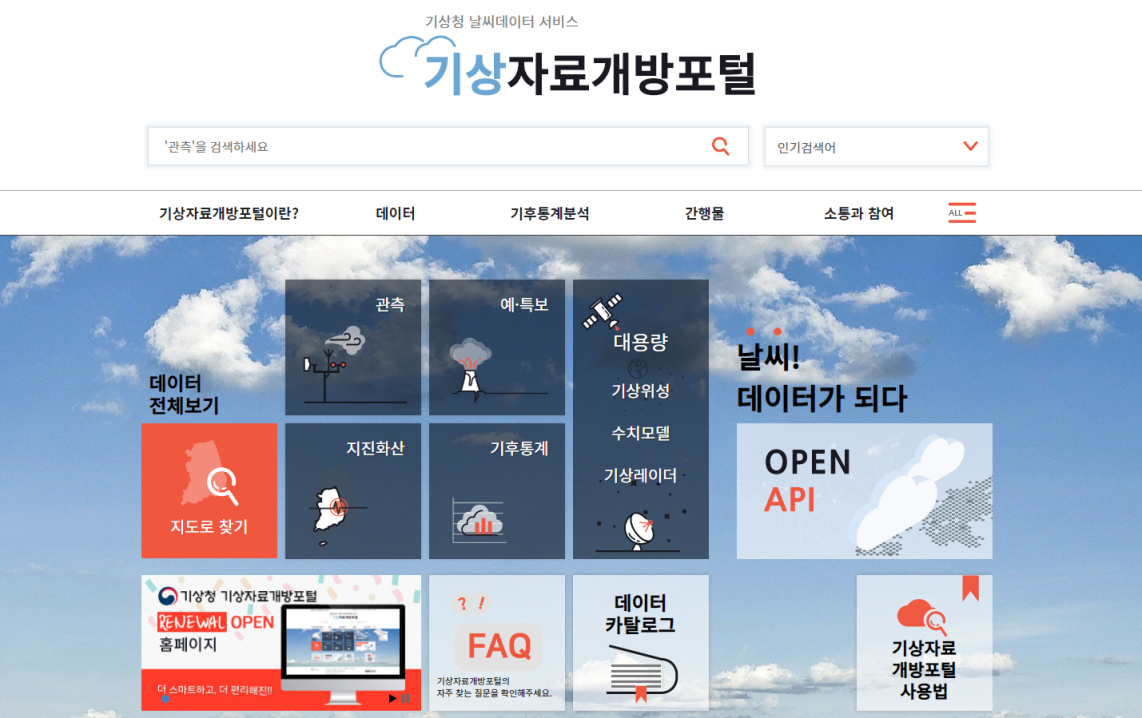


[ 그림 5 ] 전국 대기 오염 물질 자료

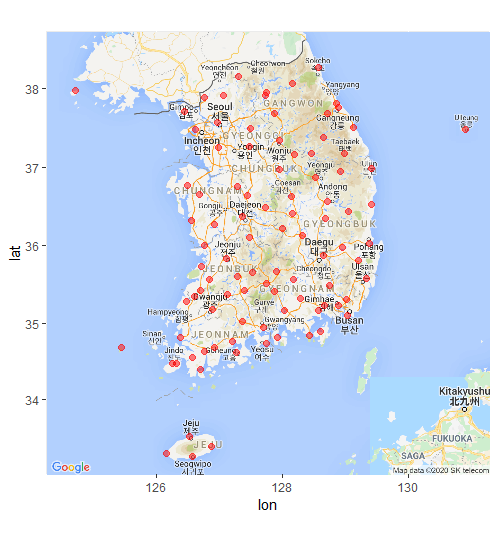
2.2 기상 데이터 전처리

기상정보는 기상청에서 제공하는 기상자료 개방 포털(http://data.kma.go.kr)에서 취득하였으며 종관기상관측 자료를 수집하였다.

종관기상관측자료(ASOS, Automated Synoptic Observing System)는 기압, 기온, 풍향, 풍속, 습도, 일사량, 일조시간 등 대기 상태를 파악하기 위해 관측하는 자료이며, 방재기상관측자료(AWS, Automatic Weather System)는 기상관측소가 없는 곳의 국지적 악기상 상황을 관측하는 곳으로 기압, 기온, 풍향, 풍속 등을 관측한다.



[ 그림 6 ] 기상자료 개방 포털 홈페이지 메인화면



[ 그림 7 ] 전국 기상 측정소 현황



[ 그림 8 ] 종 기상 관측(ASOS) 자료

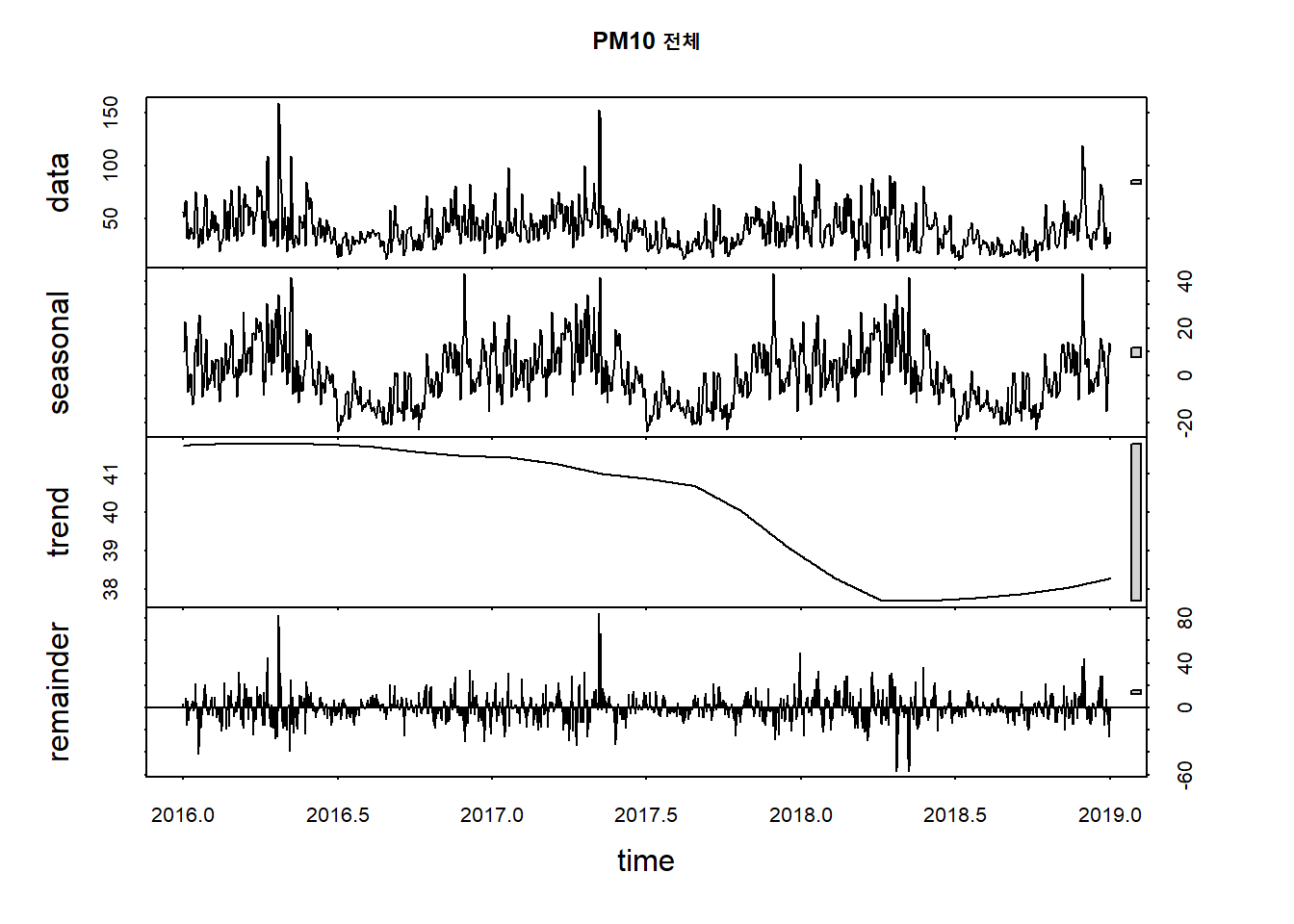
전국의 321개의 대기오염측정자료는 분기 별로 나눠진 확정자료이다. 병합 (Merge)을 통해 하나의 데이터셋으로 만들고, 시간대별 자료로 통합하였다.

2016 ~ 2018 기간의 전국 측정소의 종기상관측ASOS 데이터를 병합하여 하나의 데이터셋으로 만들고, 해당 데이터를 전국의 시/도 행정구역을 기준으로 분류하였다.

# 대기 오염 물질 분석

## 미세먼지

1.1 2016 ~ 2018년 전국 미세먼지(PM10) 일별 시계열 분석



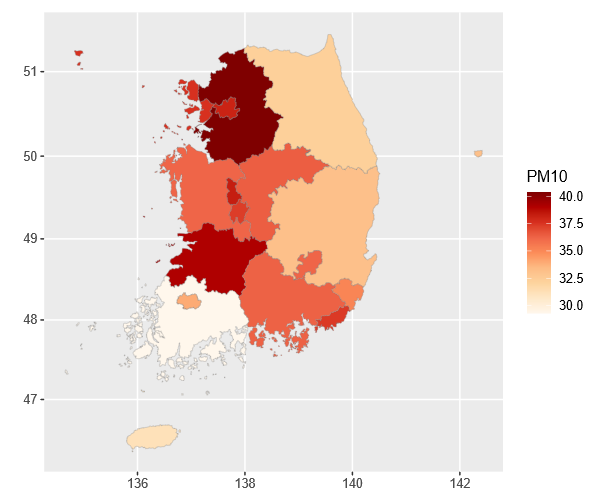
[ 그림 9 ] 2016 ~ 2018년 전국 미세먼지 시계열

2016 ~ 2018년(일별) 전국 미세먼지(PM10)를 계절성과 추세 및 나머지 요인들을 분리하여 나타낸 그래프이다.

전국 미세먼지는 4월에 최고점, 8월에는 최저점을 나타낸다. 1년을 주기로 반복되는 패턴임을 알 수 있으며 3, 4월과 11월에 유독 큰 폭의 진동을 보인다.

또한, 2016년부터 꾸준히 조금씩 감소하다가 2018년 초 이후, 미세하게 증가하는 추세임을 알 수 있다.

1.2 2016 ~ 2018년 전국 미세먼지(PM10) 지도 그래프

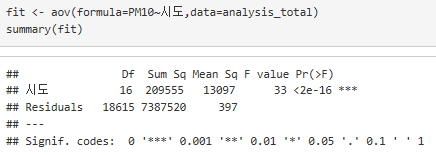


[ 그림 10 ] 2016 ~ 2018년 전국 지역별 미세먼지

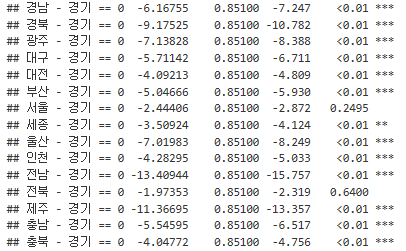
[ 그림 10 ]는 2016 ~ 2018년 전국 지역별 미세먼지 지도이다.

2016년 ~ 2018년 지역별 미세먼지(PM10)의 평균을 구하여 그래프로 나타낸 결과, 경기가 약 46으로 가장 높았고, 전남 약 32.6으로 가장 낮았다. 지역별로 평균에 차이가 있다고 볼 수 있으며, 근접한 지역끼리 어느정도 비슷한 농도를 보이는 것을 알 수 있다.

1.3 지역(시도)별 미세먼지(PM10)의 평균 차이에 대한 분산분석



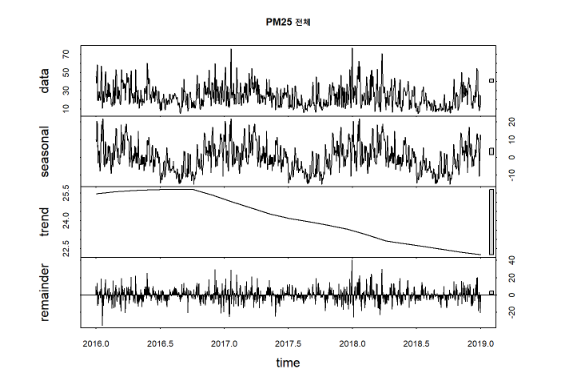
분산분석 결과, 지역에 따라 미세먼지의 평균 차이가 통계적으로 고도로 유의하다.



사후분석 결과, 경기는 주변 지역인 서울과 전북을 제외한 모든 지역에서 고도로 유의미한 차이가 있는 것으로 나타났다.

## 초미세먼지

2.1 2016 ~ 2018년 전국 초미세먼지(PM2.5) 일별 시계열 분석



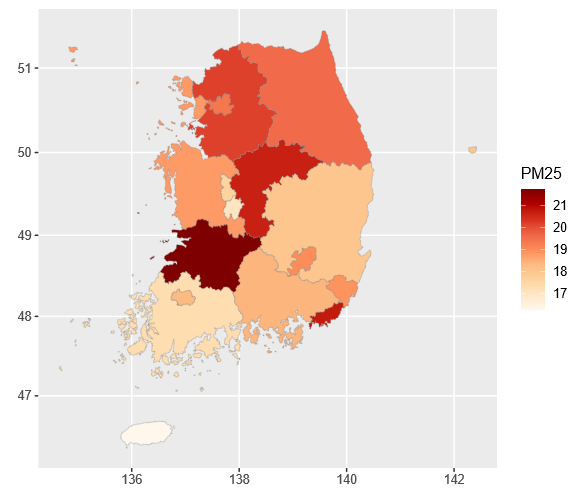
[ 그림 11 ] 2016 ~ 2018년 전국 초미세먼지 시계열

2016 ~ 2018년(일별)전국 초미세먼지(PM2.5)를 계절성과 추세 및 나머지 요인들을 분리하여 나타낸 그래프이다.

전국 초미세먼지는 1월에 최고점, 7, 8월에 최저점을 나타낸다. 1년을 주기로 반복되는 패턴임을 알 수 있으며 3, 4월에 유독 큰 폭의 진동을 보임을 알 수 있다.

또한, 2016년 상반기까지 증가하다가 시간이 지날수록 감소하는 추세이다.

2.2 2016 ~ 2018년 전국 초미세먼지(PM2.5) 지도 그래프



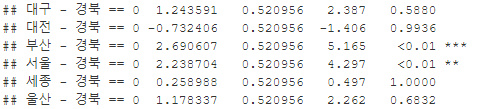
[ 그림 12 ] 2016 ~ 2018년 전국 지역별 초미세먼지

[ 그림 12 ]은 2016년 ~ 2018년 지역별 초미세먼지의 평균을 구하여 그래프로 나타낸 결과, 전북이 약 25.7로가장 높았고, 제주가 약18.4로 가장 낮았다. 지역별로 평균에 차이가 있다고 볼 수 있으며, 근접한 지역끼리 어느정도 비슷한 농도를 보이는 것을 알 수 있다.

2.3 지역(시도)별 초미세먼지의 평균 차이에 대한 분산분석



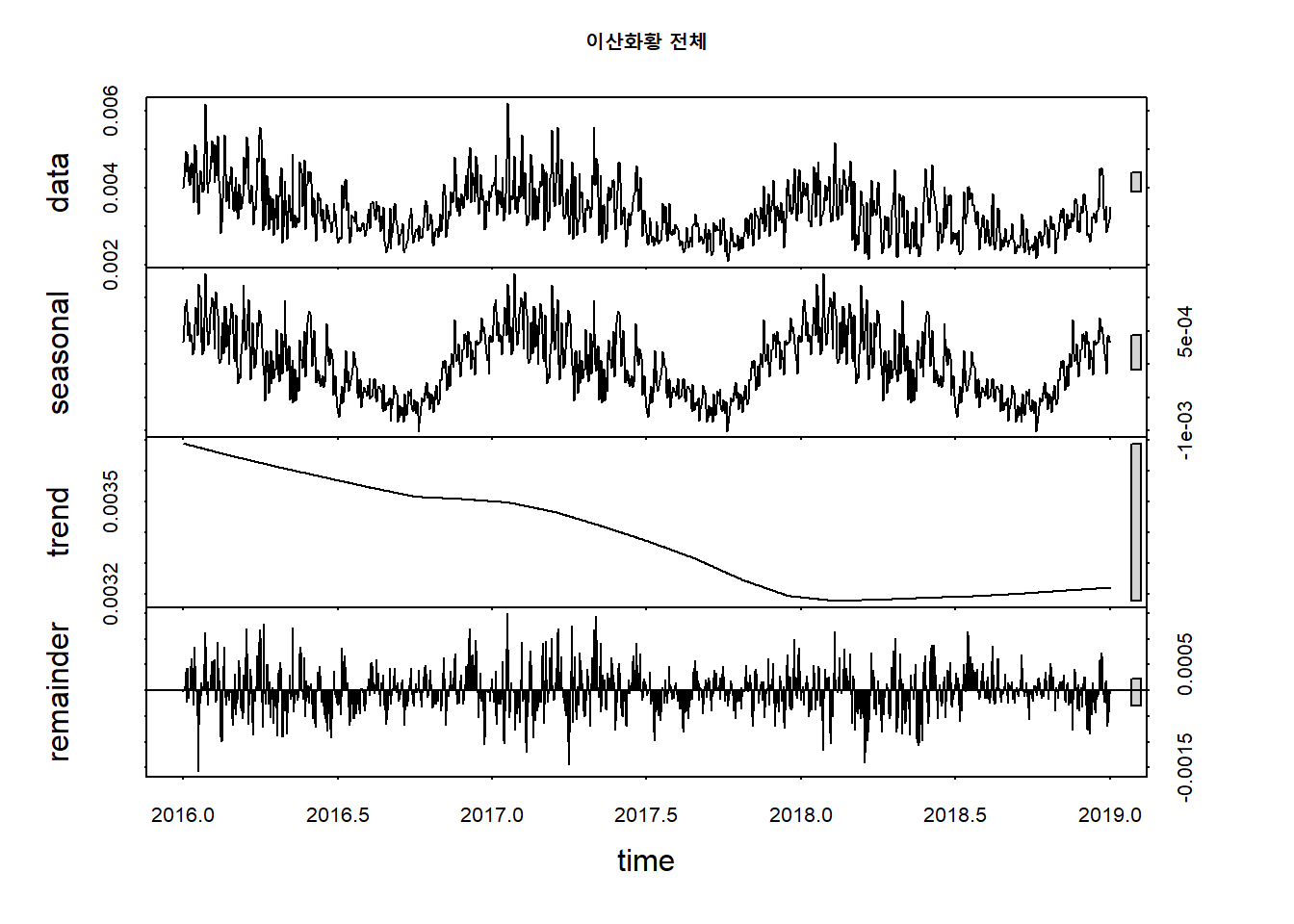
분산분석 결과, 지역에 따라 초미세먼지의 평균 차이가 통계적으로 고도로 유의하다.



사후분석 결과, 경북은 주변 지역인 대구, 대전, 세종, 울산과 유의미한 차이가 없었지만 부산, 서울과는 고도로 유의미한 차이가 있는것으로 나타났다.

## SO2

3.1 2016 ~ 2018년 전국 SO2(이산화황)일별 시계열 분석



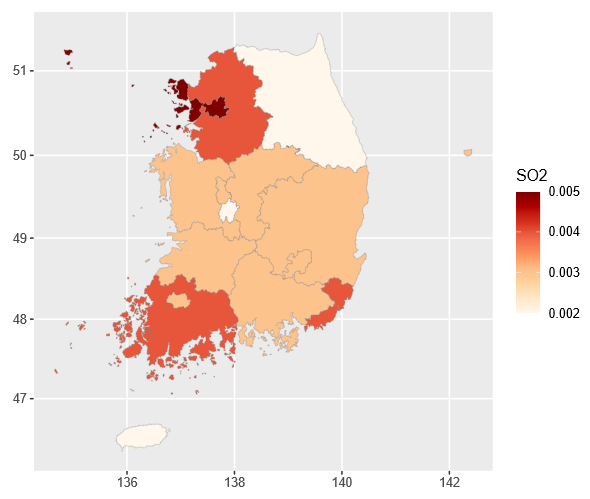
[ 그림 13 ] 2016 ~ 2018년 전국 이산화황 시계열

2016 ~ 2018년(일별)전국 SO2를 계절성과 추세 및 나머지 요인들을 분리한 그래프이다.

전국 이산화황은1월에 최고점, 7, 8월에 최저점을 나타낸다. 주로 연말에 변동폭이 감소하고, 그 외엔 변동폭이 증가함을 볼 수 있다.

또한,2016년 초 고점으로 시작하여 17년까지 하강, 2018년부터 저점으로 크게 증가하지 않는 추세임을 알 수 있다.

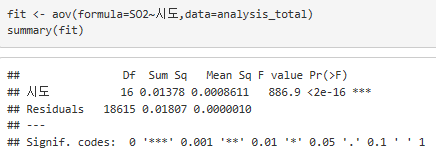
3.2 2016 ~ 2018년 전국 SO2지도 그래프



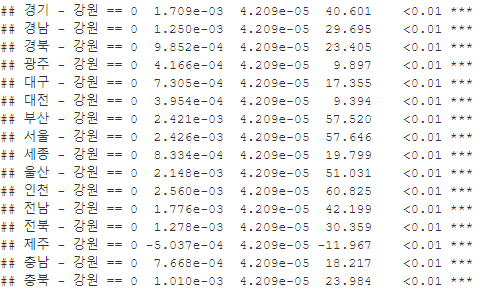
[ 그림 14 ] 2016 ~ 2018년 전국 지역별 이산화황

[ 그림 14 ] 는 2016년 ~ 2018년 지역별 SO2의 평균을 구하여 그래프로 나타낸 결과, 서울과 인천이 약 0.005로 가장 높았고, 강원과 제주가약 0.002로 가장 낮았다. 지역별로 평균에 차이가 있다고 볼 수 있으며, 근접한 지역끼리 어느정도 비슷한 수치를 보이는 것을 알 수 있다.

3.3지역(시도)별 SO2의 평균 차이에 대한 분산분석



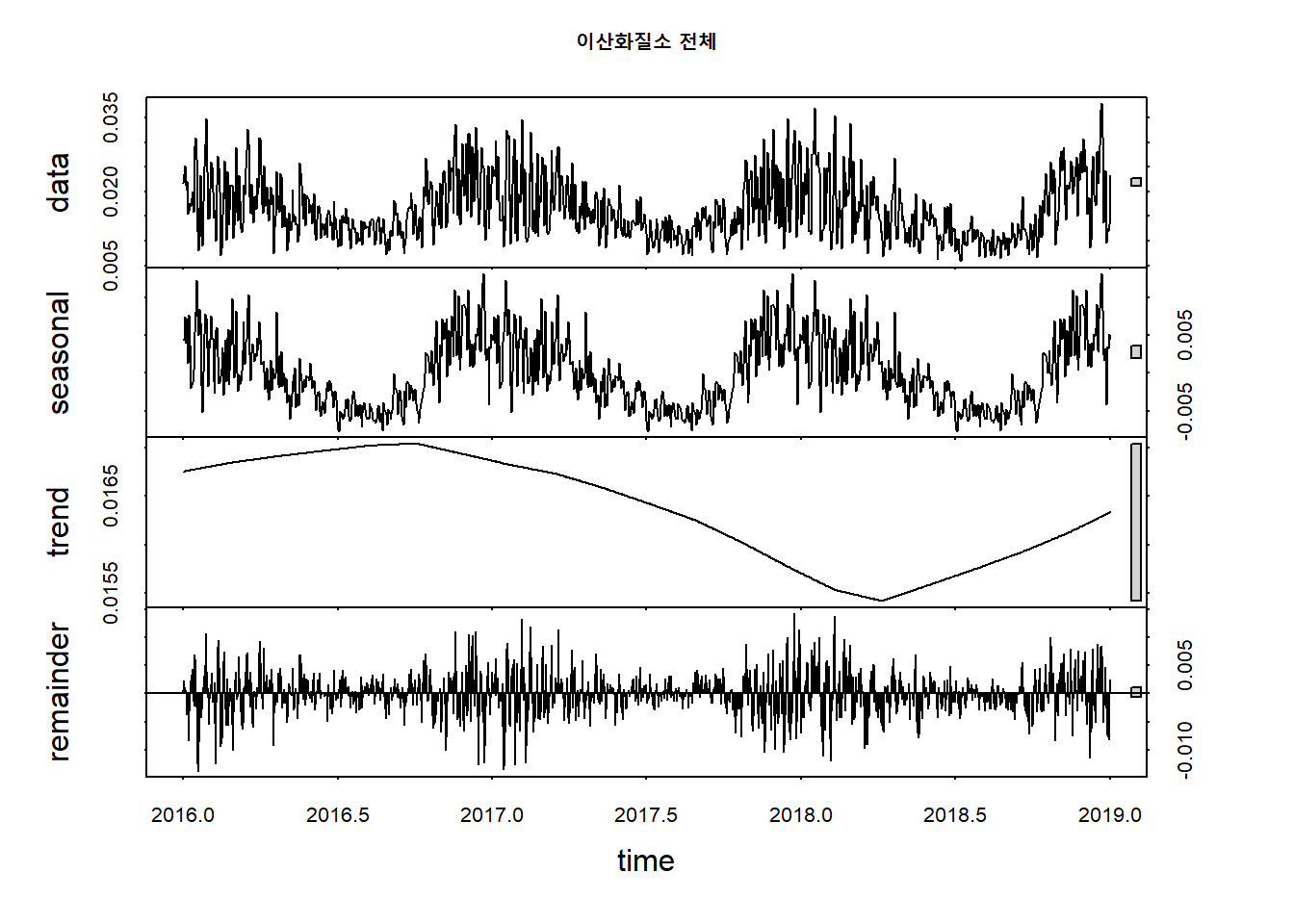
분산분석 결과, 지역에 따라 SO2의 평균 차이가 통계적으로 고도로 유의하다.



분산분석 결과, 강원은 모든 지역과 통계적으로 고도로 유의미한 차이를 보였다.

## NO2

4.1 2016 ~ 2018년 NO2(이산화질소)일별 시계열 분석



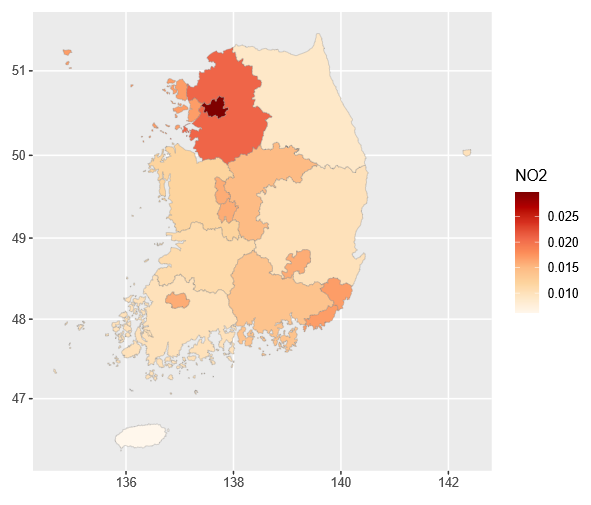
[ 그림 15 ] 2016 ~ 2018년 전국 이산화질소 시계열

2016 ~ 2018년(일별)전국 NO2를 그래프로 나타낸 그림이다.

전국 이산화질소1월에 최고점,7월에 최저점을 나타낸다.

또한, 2016년 말 이후 감소하다가 2018년 초부터 증가하는 추세이다.

4.2 2016 ~ 2018년 전국 NO2지도 그래프



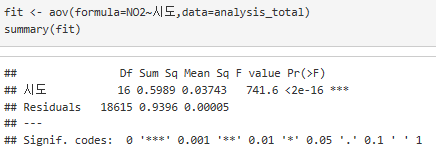
[ 그림 16 ] 2016 ~ 2018년 전국 지역별 이산화질소

2016년 ~ 2018년 지역별 NO2의 평균을 구하여 그래프로 나타낸 결과, 서울이 약 0.032로 가장 높았고, 제주가 약 0.07로 가장 낮았다. 지역별로 평균에 차이가 있다고 볼 수 있으며, 근접한 지역끼리 어느정도 비슷한 수치를 보이는 것을 알 수 있다.

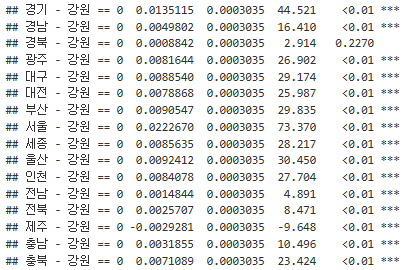
공장지대인 울산이 다른 지역에 비해 높은 양상을 보이며, 울산과 근접한 부산, 그리고 수도권인 서울과 인천이 높은 양상을 보인다.

대기오염물질(SO2, NO2, O3, CO)이 비슷한 양상을 보인다.

4.3 지역(시도)별 NO2의 평균 차이에 대한 분산분석



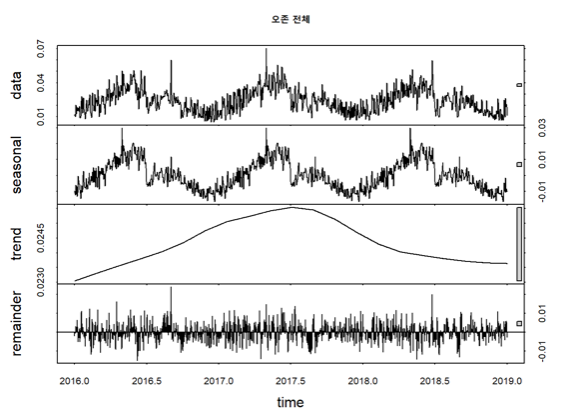
분산분석 결과, 지역에 따라 NO2의 평균 차이가 통계적으로 고도로 유의하다.



사후분석 결과, 강원은 경북을 제외한 모든 지역에서 통계적으로 고도로 유의미한 차이를 나타냈다.

## O3

5.1 2016 ~ 2018년 O3(오존)일별 시계열 분석



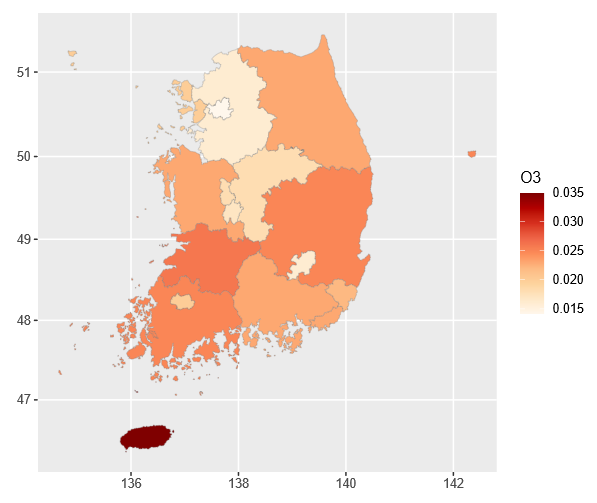
[ 그림 17 ] 2016 ~ 2018년 전국 오존 시계열

2016 ~ 2018년(일별)전국 O3를 그래프로 나타낸 그림이다.

전국 O3는 5월에 최고점,12월에 최저점을 나타낸다. 1년을 주기로 반복되는 패턴임을 알 수 있으며6월 이후 갑자기 감소하는 지점이 있다.

또한, 2017년 6월 이후 감소하는 추세임을 알 수 있다.

5.2 2016 ~ 2018년 전국 O3지도 그래프

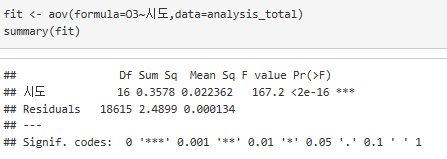


[ 그림 18 ] 2016 ~ 2018년 전국 지역별 오존

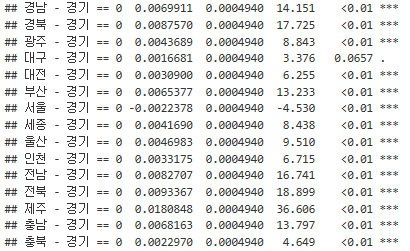
[ 그림 18 ] 은 2016년 ~ 2018년 지역별 O3의 평균을 구하여 그래프로 나타낸 결과, 제주가 약 0.037로 가장 높았고, 서울이 약 0.017로 가장 낮았다. 지역 별로 평균에 차이가 있다고 볼 수 있으며, 근접한 지역끼리 어느정도 비슷한 수치를 보이는 것을 알 수 있다.

내륙지방은 O3가 상대적으로 낮고, 바다로 둘러싸인 제주도는 상대적으로 O3 수치가 높음을 알 수 있다.

5.3지역(시도)별 O3의 평균 차이에 대한 분산분석



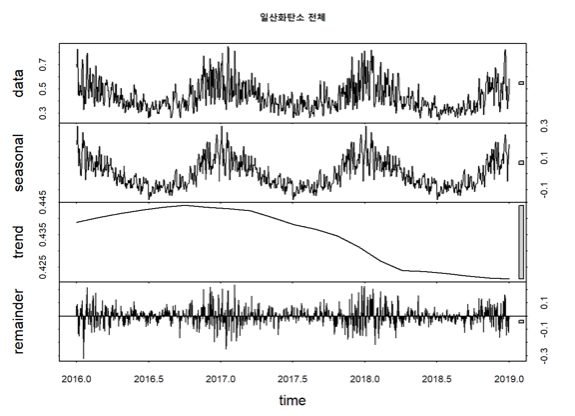
분산분석 결과, 지역에 따라 O3의평균 차이가 통계적으로 고도로 유의하다.



사후분석 결과, 경기는 대구를 제외한 모든 지역과 통계적으로 고도로 유의미한 차이를 보였다.

## CO

6.1 2016 ~ 2018년 전국 CO(일산화탄소)일별 시계열 분석



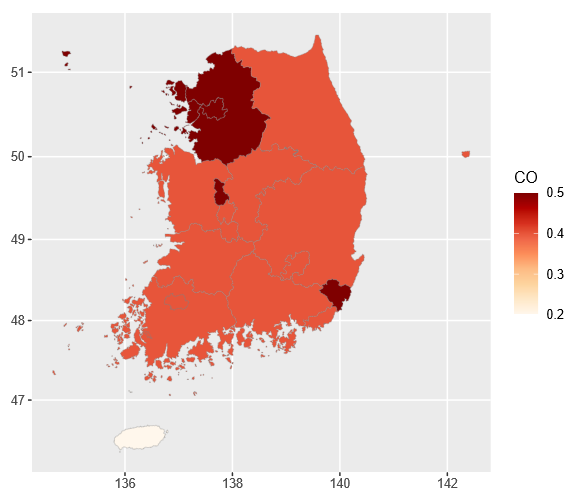
[ 그림 19 ] 2016 ~ 2018년 전국 일산화탄소 시계열

2016 ~ 2018년(일별)전국 CO를 그래프로 나타낸 그림이다.

전국 CO는7월에 최저점을 나타낸다. 1년을 주기로 반복되는 패턴임을 알 수 있으며 3, 4월에 유독 큰 폭의 진동을 보임을 알 수 있다.

또한,2017년부터 감소하여 2018년 중순으로는 감소의 폭이 줄어드는 추세이다.

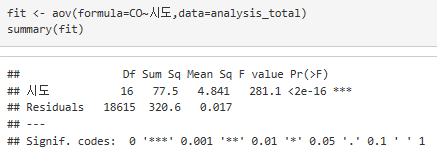
6.2 2016 ~ 2018년 전국 CO지도 그래프



[ 그림 20 ] 2016 ~ 2018년 전국 지역별 일산화탄소

[ 그림 20 ] 는 2016년 ~ 2018년 지역별 CO의 평균을 구하여 그래프로 나타낸 결과, 서울, 경기, 인천, 광주, 울산이 0.5로 가장 높았고, 제주가 약 0.3으로 가장 낮았다. 지역 별로 농도에 차이가 있다고 볼 수 있으며, 근접한 지역끼리 어느정도 비슷한 수치를 보이는 것을 알 수 있다.

6.3지역(시도)별 CO의 평균 차이에 대한 분산분석



분산분석 결과, 지역에 따라 CO의평균 차이가 통계적으로 고도로 유의하다.

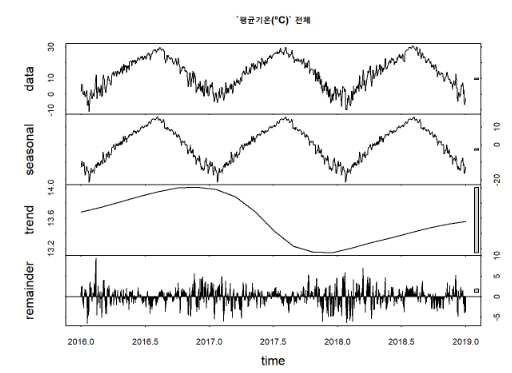


사후분석 결과, 강원도는 경남, 대구를 제외한 모든 지역에서 통계적으로 고도로 유의미한 차이를 보였다.

# 기상 분석

## 기온

* 1. 2016 ~ 2018년 전국 일별 시계열 분석

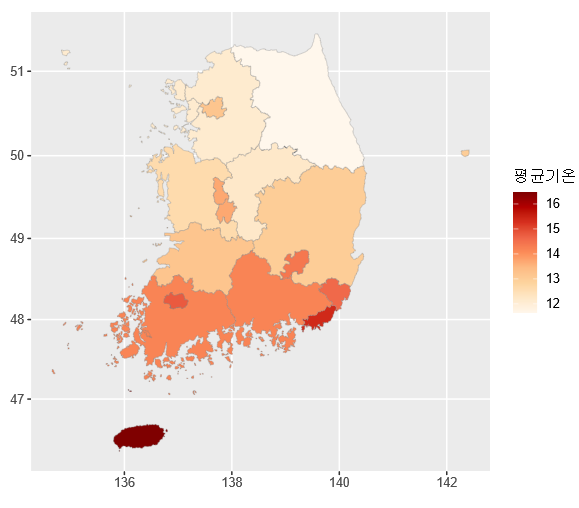


[ 그림 21 ] 2016 ~ 2018년 전국 평균기온 시계열

전국 기온 데이터 중 2016년 ~ 2018년까지 3 년 간 일 평균 기온의 시계열 패턴을 분석하였다. STL(Seasonal and Trend decomposition using Loess) 방법으로 추세 성분, 계절 성분, 불규칙 성분으로 나누어 각각의 패턴을 분석하였다.

전국 평균 기온의 계절요인은 연내 여름~겨울 계절성에 따른 기온변동에 기인하며, 3년간 전국 평균기온의 추세요인을 보면 2016~2017년까지 상승하다가 2017년 연중에 감소하여 2018년에도 낮은 추세를 보인다.

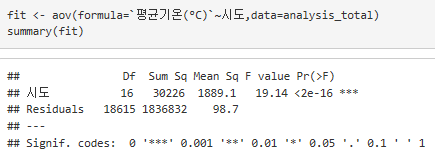
* 1. 2016 ~ 2018년 전국 평균기온 지도



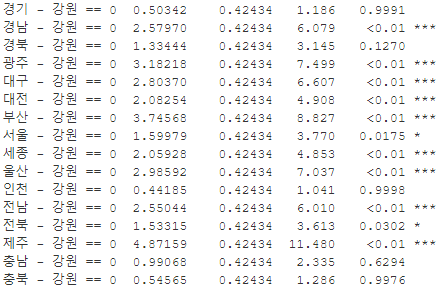
[ 그림 22 ] 2016 ~ 2018년 전국 지역별 평균기온

[ 그림 22 ] 는 2016년 ~ 2018년 지역별 평균 기온을 구하여 그래프로 나타낸 결과, 제주가 16.5로 가장 높았고, 강원이 11.6으로 가장 낮았다. 지역별로 기온에 차이가 있다고 볼 수 있으며, 근접한 지역끼리 어느정도 비슷한 수치를 보이는 것을 알 수 있다.

1.3 지역(시도)별 평균기온의 차이에 대한 분산분석



분산분석 결과, 지역에 따라 평균기온의 차이가 통계적으로 고도로 유의하다.

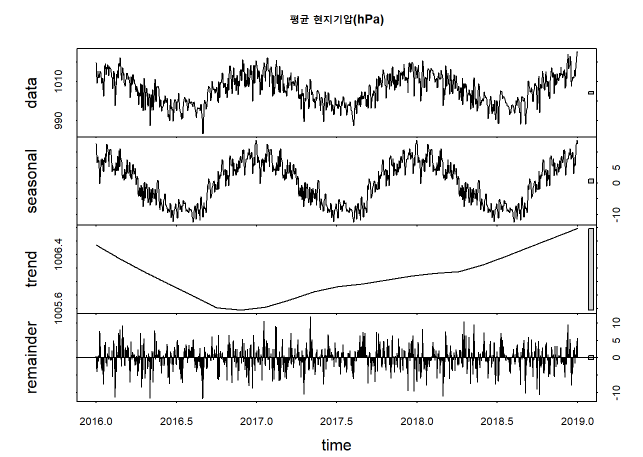


사후분석 결과, 평균기온은 강원이 가장 낮은 것으로 나타났으며, 경기, 경북, 충북 등 중부지방과 통계적으로 유의미한 차이를 보이지 않았으며, 남부지방과는 고도로 유의미한 차이를 보였다. 서울은 중부지방임에도 강원도와 통계적으로 유의미한 차이를 나타냈는데, 이는 서울의 인구밀집과 관련하여 열섬 현상이 나타나기 때문인 것으로 추측된다.

## 기압(평균현지기압)

전국 기상 관측 데이터셋 중 평균 현지기압(hPa)으로 측정된 기상데이터를 연도별로 분류한 후 월 평균 기압(hPa)을 시각화하여 분석하였다.

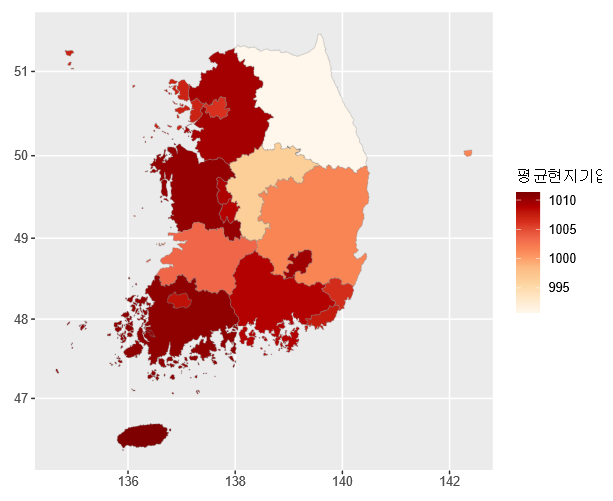
* 1. 2016 ~ 2018년 전국 평균 현지기압 시계열 분석



[ 그림 23 ] 2016 ~ 2018년 전국 현지기압 시계열

전국 평균 현지기압(hPa)의 계절요인을 보면 겨울 때 최고점을 찍으며, 추세를 보면 연도가 흐를수록 올라가는 추세이다. 추세적으로 2016년도 까지는 감소하다가 2017년도부터 증가 추세를 보인다.

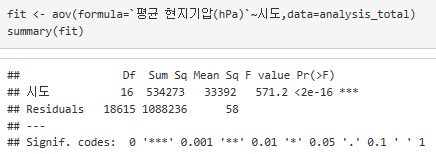
* 1. 2016 ~ 2018년 전국 평균 현지기압 지도



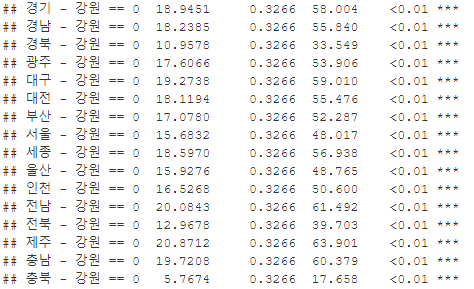
[ 그림 24 ] 2016 ~ 2018년 전국 지역별 평균 현지기압

2016년 ~ 2018년 지역별 평균 현지기압을 구하여 그래프로 나타낸 결과, 제주가 1011.5로 가장 높았고, 강원이 990.6으로 가장 낮았다. 지역별로 평균 현지기압에 차이가 있다고 볼 수 있으며, 근접한 지역끼리 어느정도 비슷한 수치를 보이는 것을 알 수 있다.

2.3 지역(시도)별 평균 현지기압의 차이에 대한 분산분석



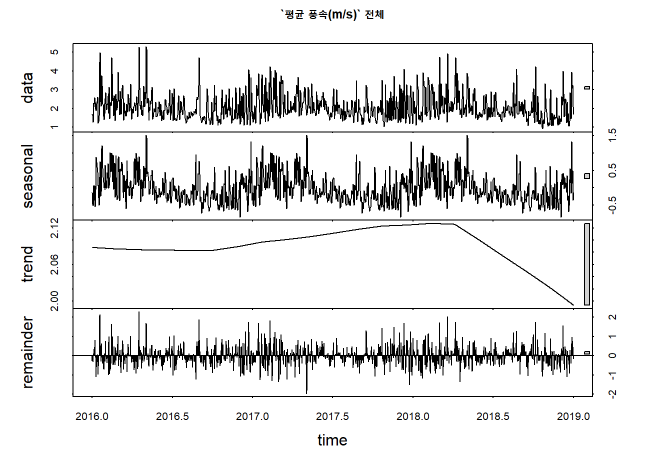
분산분석 결과, 지역에 따라 평균 현지기압의 차이가 통계적으로 고도로 유의하다.



사후 분석 결과, 강원의 평균현지기압이 가장 낮았으며, 통계적으로 모든 지역과 고도로 유의미한 차이를 보였다.

## 풍속

* 1. 2016 ~ 2018년 전국 평균 풍속 시계열 분석

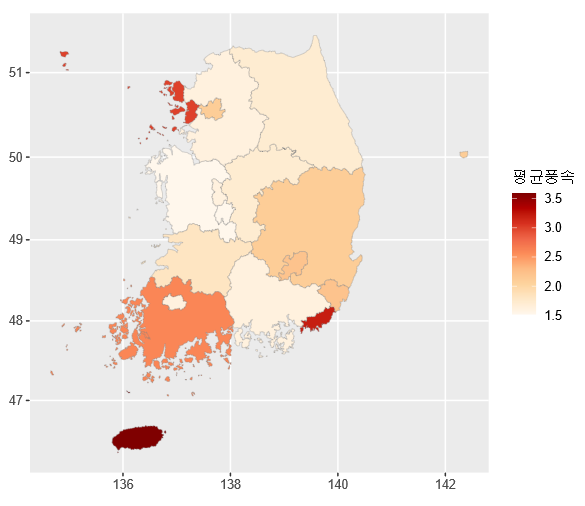


[ 그림 25 ] 2016 ~ 2018년 전국 평균 풍속 시계열

1년을 주기로 반복되며, 12월에 최고점, 7~8월에 최저점을 기록한다. 3~4월에 풍속 유난히 높게 나타나는 것을 볼 수 있는데 이는 꽃샘추위와 관계가 있는 것으로 예측된다.

전체적인 추세는 2018년 초까지 완만하게 증가하다 그 이후 하락하는 경향을 보인다.

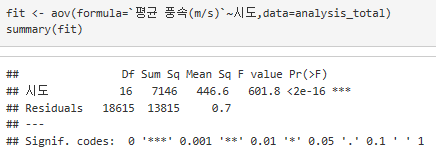
* 1. 2016 ~ 2018년 전국 평균 풍속 지도



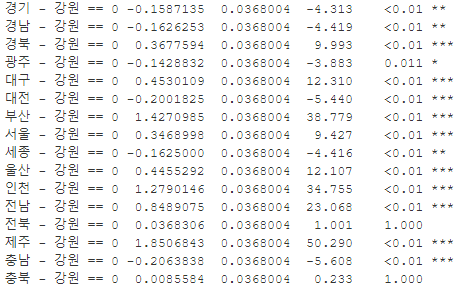
[ 그림 26 ] 2016 ~ 2018년 전국 지역별 평균 풍속

2016년 ~ 2018년 지역별 평균 풍속을 구하여 그래프로 나타낸 결과, 제주가 3.6으로 가장 높았고, 충남이 1.5로 가장 낮았다. 지역별로 평균 풍속에 차이가 있다고 볼 수 있으며, 근접한 지역끼리 어느정도 비슷한 수치를 보이는 것을 알 수 있다.

* 1. 지역(시도)별 평균 풍속의 차이에 대한 분산분석



분산분석 결과, 지역에 따라 평균 풍속의 차이가 통계적으로 고도로 유의하다.

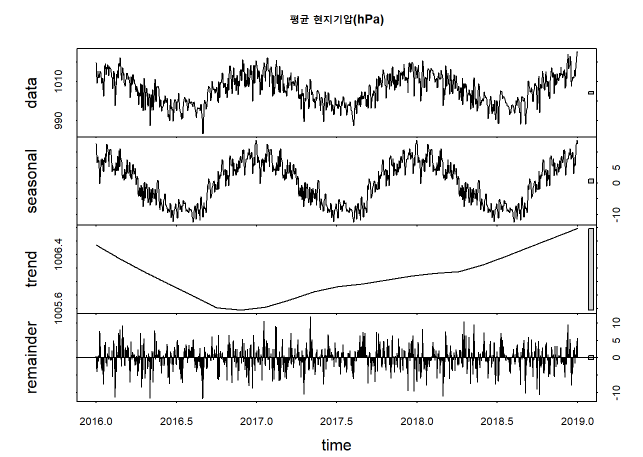


분산분석 결과, 강원은 전북, 충북과는 유의미한 차이가 없는 것으로 나타났다.

## 풍향

풍향의 분포는 지역별 국지적인 영향으로 특정한 패턴이 없는 것으로 추정이 되며 전라북도 지역은 1분기 연도별 200도 내외의 풍향으로 나타나지만 큰 특징은 나타나지 않는다. 지상부의 풍향이 특별한 패턴을 보이지 않는 것은 낮은 고도의 기류는 변화 무쌍하며 주위의 건물 및 지형상황에 따라 흐름의 변동이 다르기 때문으로 판단된다.

* 1. 2016 ~ 2018년 전국 연도별 일별 시계열 분석



[ 그림 27 ] 2016 ~ 2018년 전국 평균 현지기압 시계열

전국 평균 현지기압(hPa)의 계절요인을 보면 겨울 때 최고점을 찍으며, 추세를 보면 연도가 흐를수록 올라가는 추세이다. 추세적으로 2016년도 까지는 감소 추세이다 2017년도부터 증가 추세를 보인다.

* 1. 2016 ~ 2018년 전국 풍향 빈도 수 지도

2016년 ~ 2018년 지역별 평균 기온을 구하여 그래프로 나타낸 결과, 제주가 16.5로 가장 높았고, 강원이 11.6으로 가장 낮았다. 지역별로 기온에 차이가 있다고 볼 수 있으며, 근접한 지역끼리 어느정도 비슷한 수치를 보이는 것을 알 수 있다.

4.3 지역(시도)별 풍향 빈도수 차이에 대한 분산분석

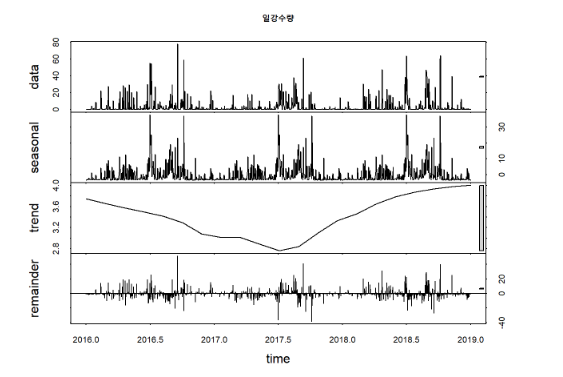
분산분석 결과, 지역에 따라 평균기온의 차이가 통계적으로 고도로 유의하다.

## 강수

## 

전국 기상 관측 데이터셋 중 강수량 데이터를 연도별로 분류하여 월 평균 강수량(mm)을 시각화하여 분석하였다.

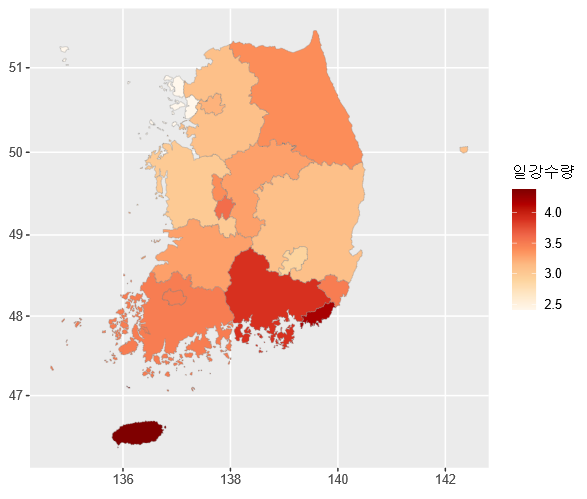
5.1 2016 ~ 2018년 전국 일강수량 시계열 분석



[ 그림 28 ] 2016 ~ 2018년 전국 일강수량 시계열

강수량은 여름에서 초가을까지 최고조를 찍는 것으로 보인다.2017년까지 감소추세를 보이다 2017년 여름이후 증가추세를 보인다.

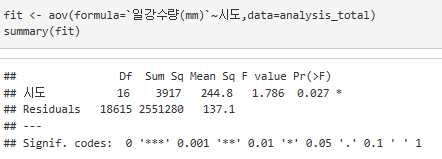
5.2 2016 ~ 2018년 전국 일강수량 지도



[ 그림 29 ] 2016 ~ 2018년 전국 지역별 일강수량

2016년 ~ 2018년 지역별 평균 기온을 구하여 그래프로 나타낸 결과, 제주가 16.5로 가장 높았고, 강원이 11.6으로 가장 낮았다. 지역별로 기온에 차이가 있다고 볼 수 있으며, 근접한 지역끼리 어느정도 비슷한 수치를 보이는 것을 알 수 있다.

5.1.3 지역(시도)별 일강수량의 차이에 대한 분산분석



분산분석 결과, 지역에 따라 일강수량의 차이가 통계적으로 유의하다.

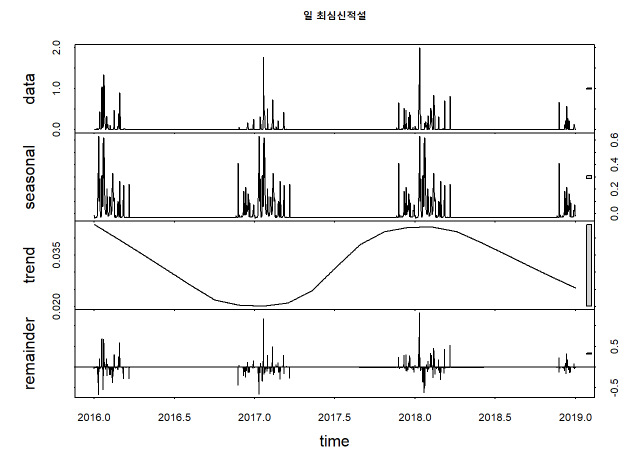




사후분석 결과, 일강수량은 대부분의 지역에서 통계적으로 차이가 없는 것으로 나타났지만, 인천과 제주는 고도로 유의미한 차이를 보였고, 인천과 부산은 유의미한 차이를 보냈다.

## 강설(최심신적설)

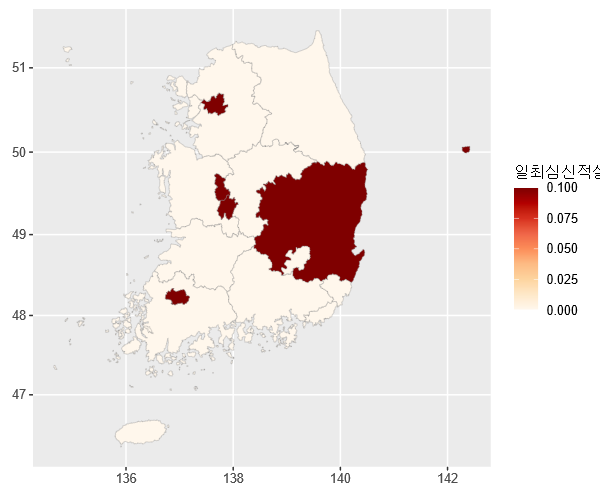
* 1. 2016 ~ 2018년 전국 강설(최심신적설) 시계열 분석



[ 그림 32 ] 2016 ~ 2018년 전국 일 최심신적설 시계열

일 최심신적설은 주로 1, 2월에 최고점을 찍는다.

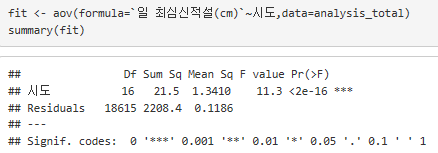
* 1. 2016 ~ 2018년 전국 일 최심신적설 지도



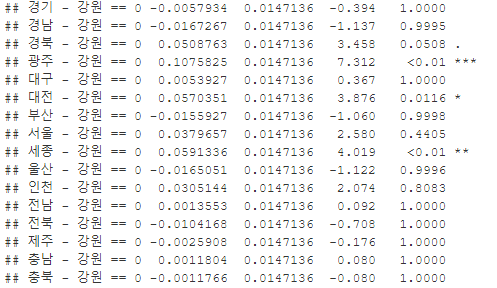
[ 그림 33 ] 2016 ~ 2018년 전국 지역별 일 최심신적설

2016년 ~ 2018년 지역별로 일 최심신적설을 그래프로 나타낸 결과, 경북, 서울, 세종, 대전, 광주가 0.1로 가장 높았다. 지역별로 일 최심신적설에 차이가 있다고 볼 수 있으며, 근접한 지역끼리 어느정도 비슷한 수치를 보이는 것을 알 수 있다.

6.3 지역(시도)별 일 최심신적설 차이에 대한 분산분석



분산분석 결과, 지역에 따라 일 최심신적설의 차이가 통계적으로 고도로 유의하다.



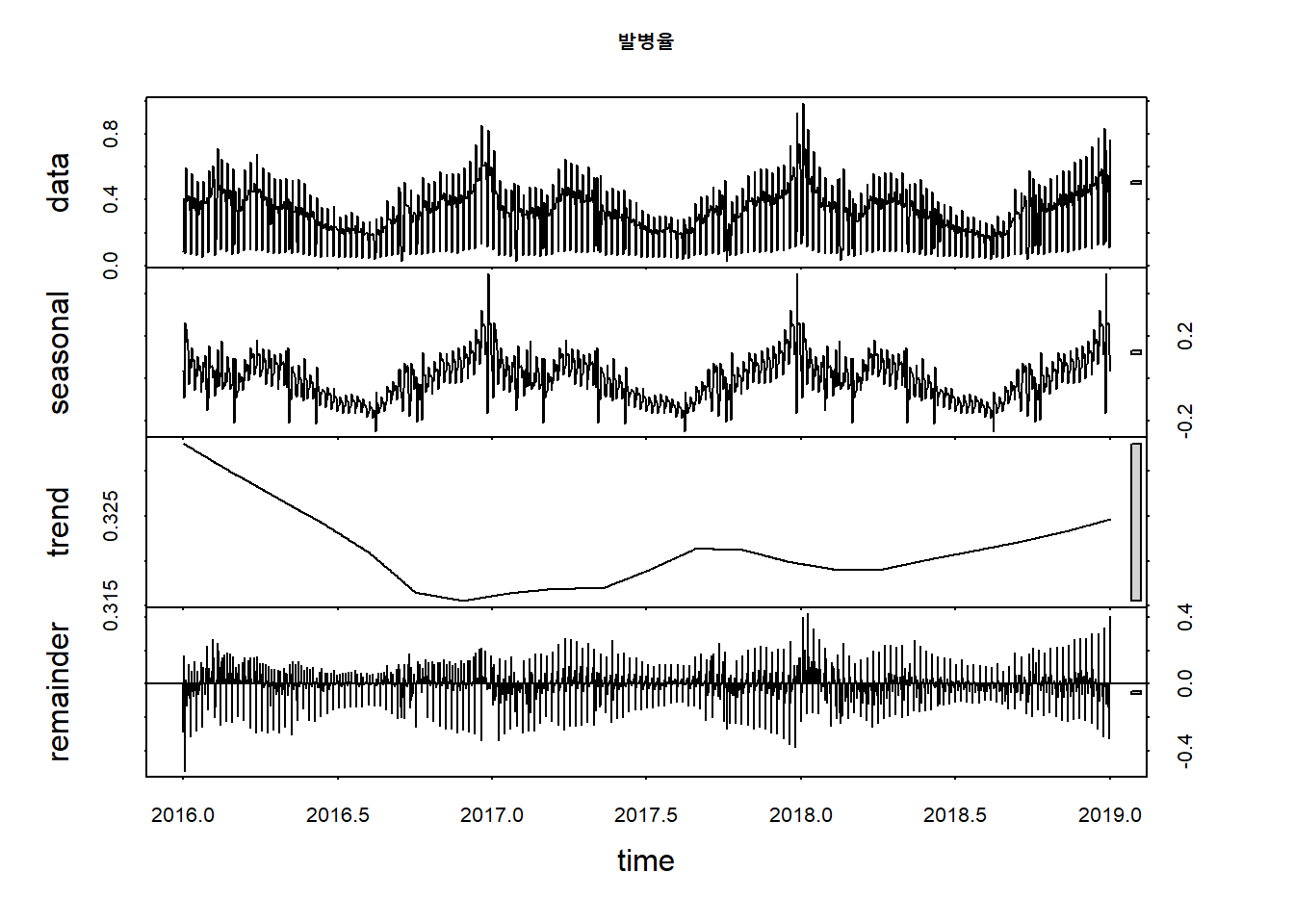
사후분석 결과, 강원은 광주, 세종과 고도로 유의미한 차이를 보였으며, 대전과 유의미한 차이를 나타냈다.

# 진료내역 분석

## 발병률 분석

환자 의료 데이터 중 호흡기 질환 코드인 J 코드를 추출하여 분석하였다.각 지역,일별로 호흡기 질환 진료건수를 발병건수로 설정하고,각 지역별 인구수로 나눠주어 발병률 파생변수를 생성했다.

1.1 2016 ~ 2018년 전국 발병률일별 시계열 분석

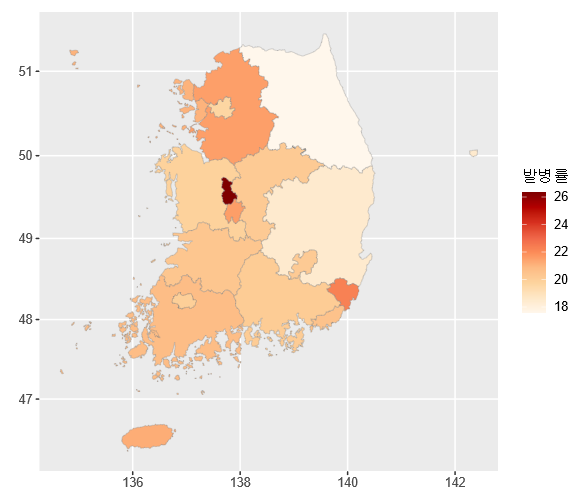


[ 그림 34 ] 2016 ~ 2018년 전국 발병률 시계열

호흡기 질환 발병률은 12, 1월에 최고점을 찍고, 감소하다가 3, 4월에 급증함을 알 수 있다. 그 후,서서히 감소하여 7, 8월에 최저점을 찍는다.

미세먼지가 대부분 3, 4월에 증가하는 추세임을 고려해봤을 때, 발병률에 미세먼지가 어느정도 영향을 끼친다고 생각할 수 있다.

* 1. 2016 ~ 2018년 전국 발병률 지도 그래프



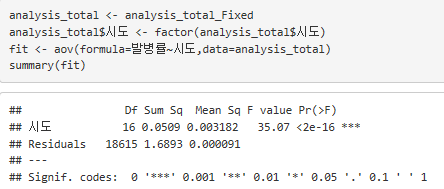
[ 그림 35 ] 2016 ~ 2018년 전국 지역별 발병률

2016년 ~ 2018년 지역별 발병률의 평균을 구하여 그래프로 나타냈다. 발병률은 수치상 알아보기 쉽게 발병건수 \* 10,000으로 계산하여, 인구 10,000명 당 발병건수로 나타냈다.

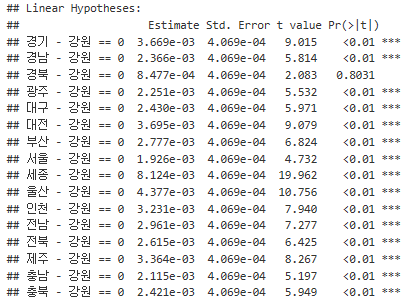
세종이 약 26.402로 가장 높았고, 강원이 약 17.498로 가장 낮았다. 지역별로 평균에 차이가 있다고 볼 수 있으며, 근접한 지역끼리 어느정도 비슷한 수치를 보이는 것을 알 수 있다.

주로, 동쪽 지역은 발병률이상대적으로 낮고, 서쪽은 평균 이상임을 볼 수 있다.

1.3 지역(시도)별 발병률의 평균 차이에 대한 분산분석



분산분석 결과, 지역에 따라 발병률의 평균 차이가 통계적으로 고도로 유의하다.

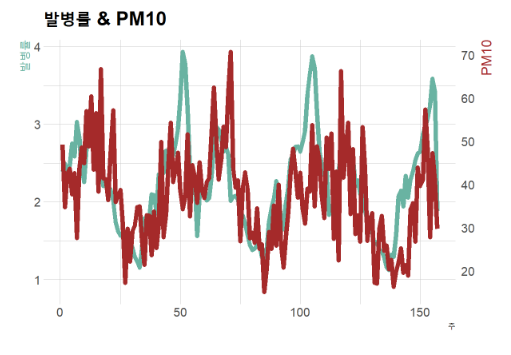


사후분석 결과, 강원은 경북을 제외한 모든 지역에서 고도로 유의미한 차이가 있는 것으로 나타났다.

## 기상요인, 대기오염물질에 따른 발병률 분석

2.1 2016 ~ 2018년 전국 오염물질과 발병률 간 패턴 분석

2.1.1 PM10



[ 그림 36 ] 주별 발병률과 미세먼지 패턴

3, 4월에 발병률이 급증하는 구간과PM10이 최대치를 기록하는 구간이 일치하는 것을 확인했음.

발병률이 최대치를 기록하는 12월을 제외하면 패턴이 비교적으로 일치하는 것을 볼 수 있음. 이에 따라 발병률과 PM10이 연관성이 있을 것이라 추측함.

2.1.2 PM2.5



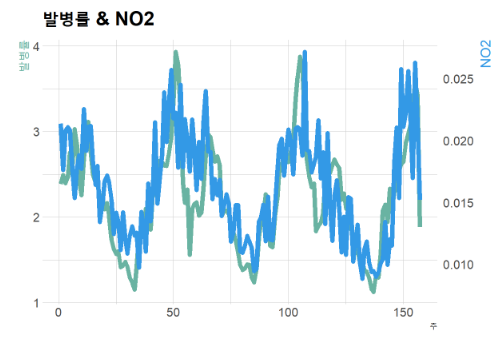
[ 그림 37 ] 주별 발병률과 초미세먼지 패턴

PM10과 비슷함을 확인.3, 4월에 발병률이 급증하는 구간과PM2.5가 최대치를 기록하는 구간이 일치하는 것을 확인했음.

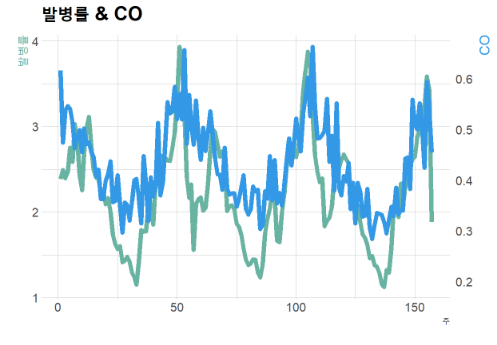
발병률이 최대치를 기록하는 12월을 제외하면 패턴이 비교적으로 일치하는 것을 볼 수 있음.

이에 따라 발병률과 PM2.5이 연관성이 있을 것이라 추측함.

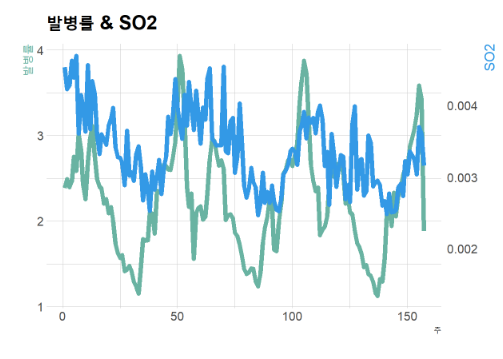
2.1.3 대기오염물질 (SO2, NO2, CO)



[ 그림 38 ] 주별 발병률과 NO2 패턴



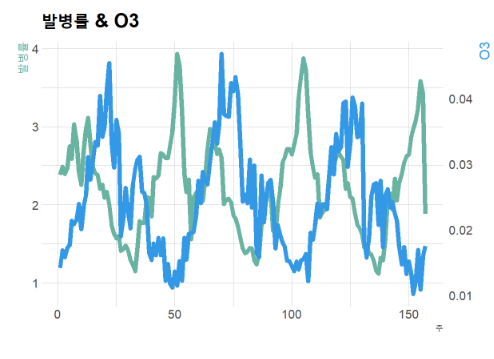
[ 그림 39 ] 주별 발병률과 CO 패턴



[ 그림 40 ] 주별 발병률과 SO2 패턴

SO2, NO2, CO는 서로 비슷한 패턴을 보이고, 발병률과의 관계에 있어서 전체적으로 비슷한 패턴을 나타내지만 1 ~ 4월까지 발병률이 급감하다가 급증하는 부분은 설명해주지 못하는 것으로 보인다.

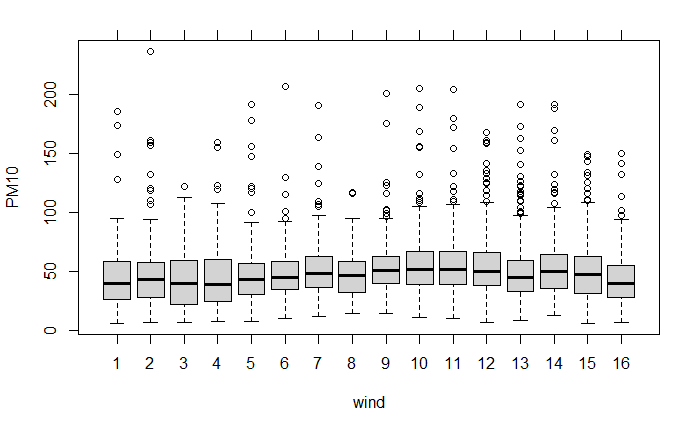
* + 1. O3



[ 그림 41 ] 주별 발병률과 O3 패턴

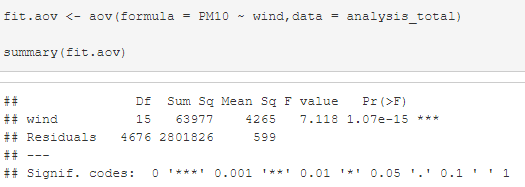
O3는 지연효과를 적용했을 때, 발병률과 가장 비슷한 패턴을 볼 수 있을 것으로 추측된다.

2.2.1 풍향 별 미세먼지 분석

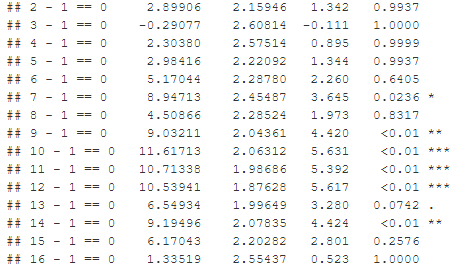


[ 그림 42 ] 풍향 별 미세먼지

PM10의 농도는 서풍이 불 때 가장 높았으며 다른 방향에서 바람이 불 때 보다 특이하게 높은 수치를 기록하는 경우가 많았다.



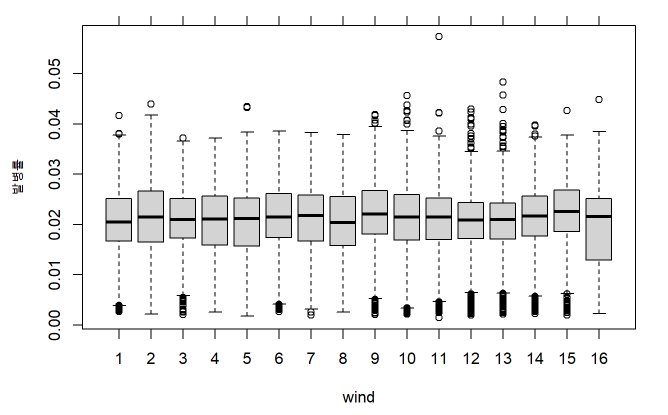
분산분석 결과, 풍향 별 미세먼지의 차이는 고도로 유의미한 것으로 나타났다.



사후분석 결과를 보면, 서풍(10~14)이 부는 경우와 나머지 풍향의 미세먼지의 평균의 차이가 고도로 유의미하게 나타나는 것을 확인할 수 있다.

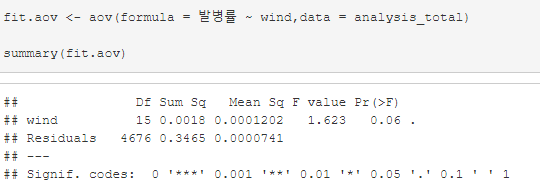
따라서, 3,4월에 미세먼지가 크게 늘어나는 이유는 서풍의 영향이 클 것으로 예상된다.

2.2.2 풍향 별 발병률 분석



[ 그림 43 ] 풍향 별 발병률

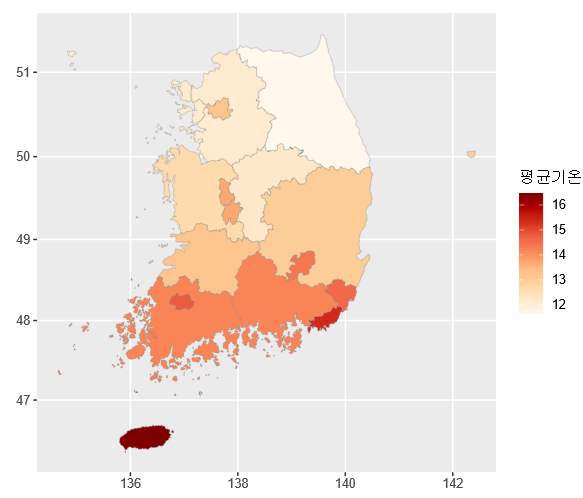
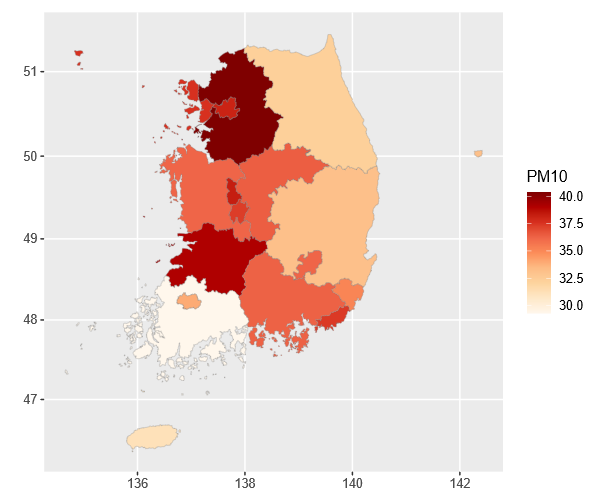
풍향 별 발병률은 크게 차이가 나지 않는 것으로 보인다.



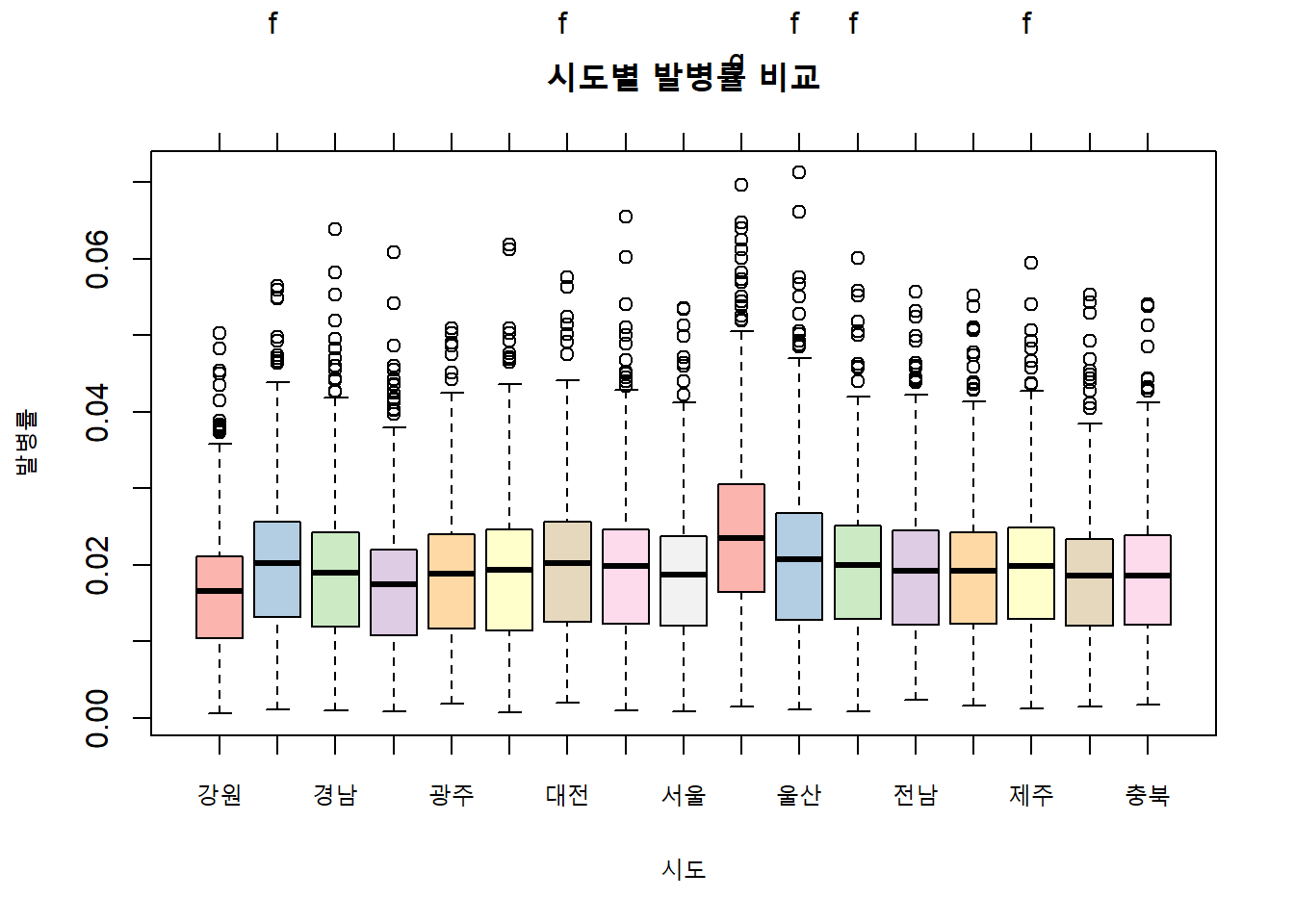
분산분석 결과, 풍향 별 미세먼지의 차이는 고도로 유의미한 것으로 나타났지만, 풍향 별 발병률의 차이는 유의미한 차이를 보이지 않는 것으로 보아 미세먼지가 발병률에 미치는 영향이 없거나 적을 것으로 예상된다.

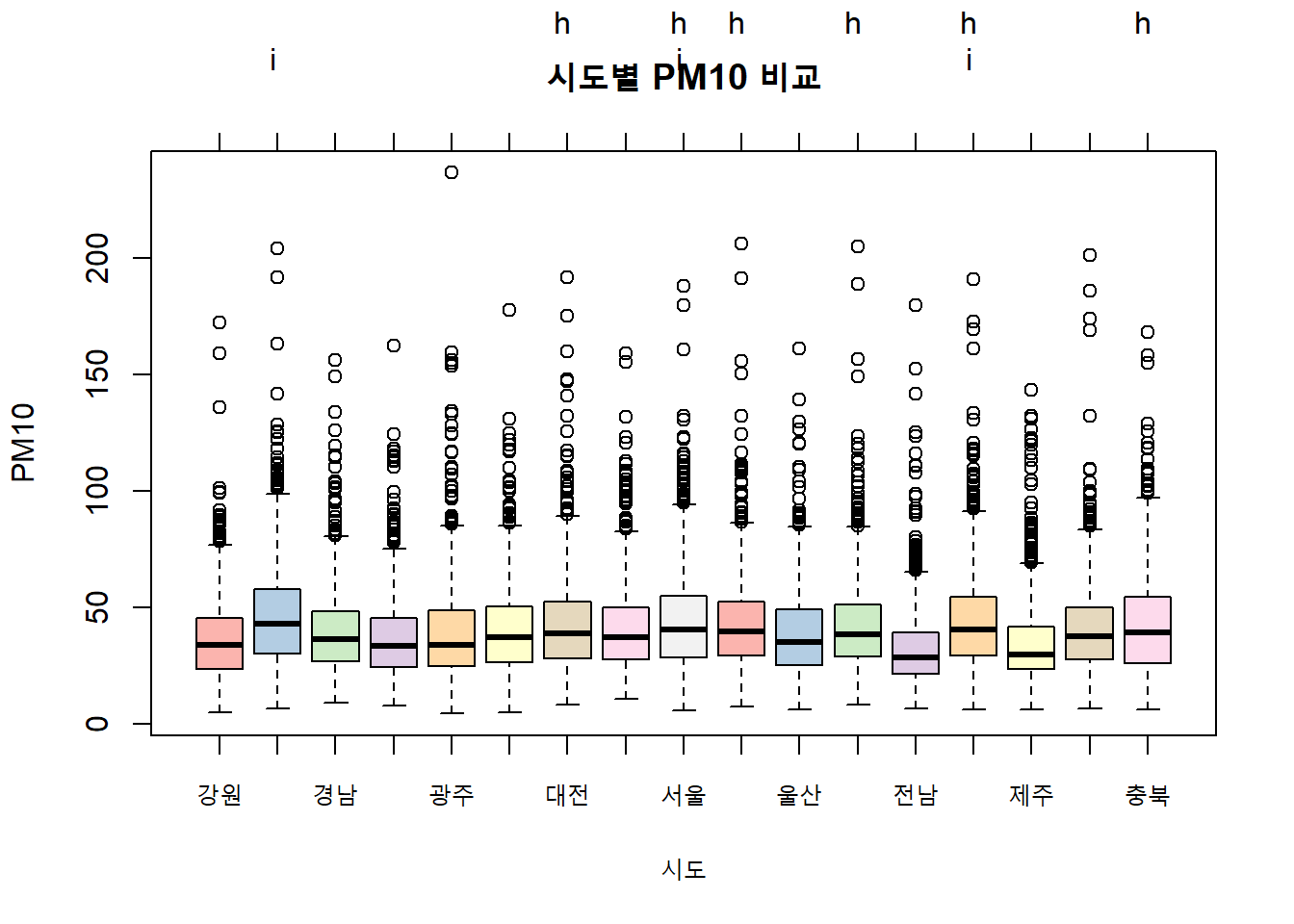
# 군집분석

## 지역별 요인 비교



[ 그림 43 ] 지역별 미세먼지, 평균기온

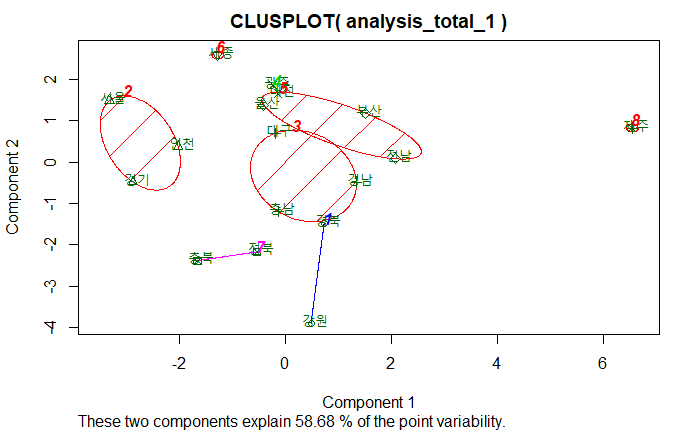
[ 그림 44 ] 시도별 발병률 비교



[ 그림 45 ] 시도별 미세먼지 비교

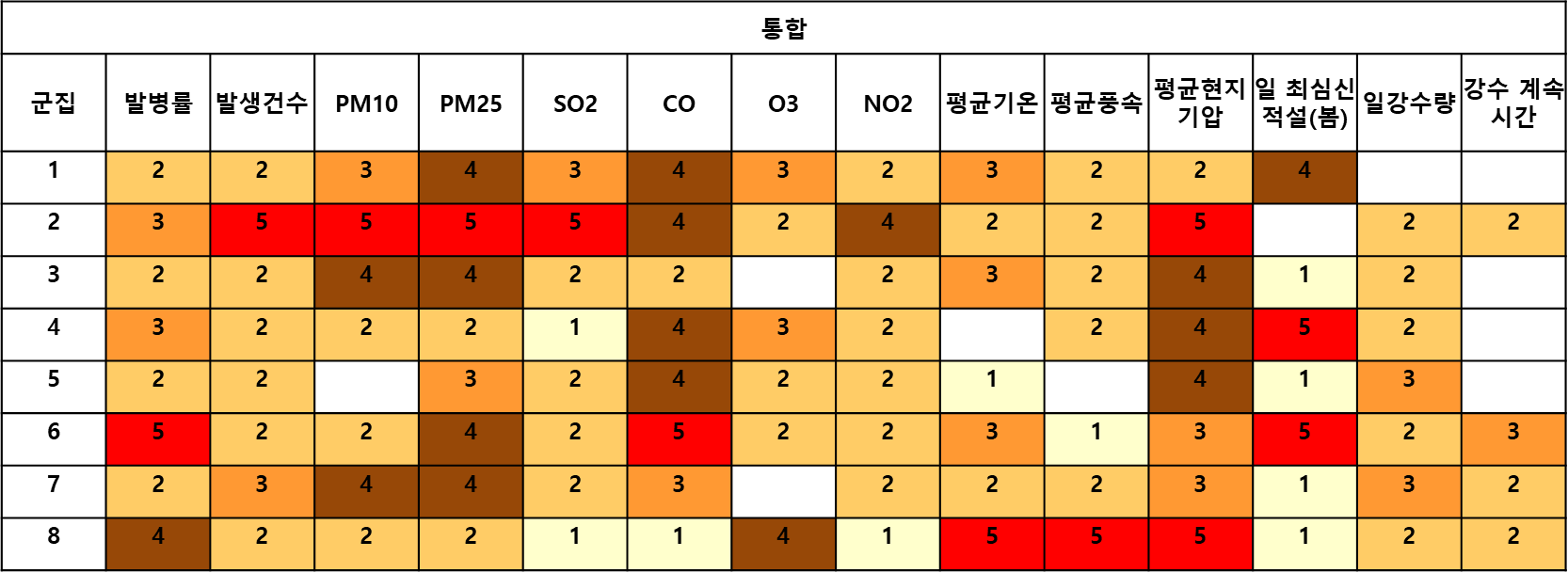
근접한 지역끼리 각 요인들이 비슷한 수치임을 알 수 있음.따라서,군집화를 통해 근접한 지역을 묶어서 각 군집별 특성을 비교하려 함.

## 군집화



[ 그림 46 ] 클러스터링

요인별 분석 결과,세종에서 발병률이 특이하게 높은 것을 확인했고,이에 따라 세종이 따로 분리되도록 k-means 알고리즘을 사용하여 군집화 한 결과, k = 8 일 때,가장 적절한 것으로 판단되었다. 아래 그림은 각 군집별 특성을 비교한 표이다.

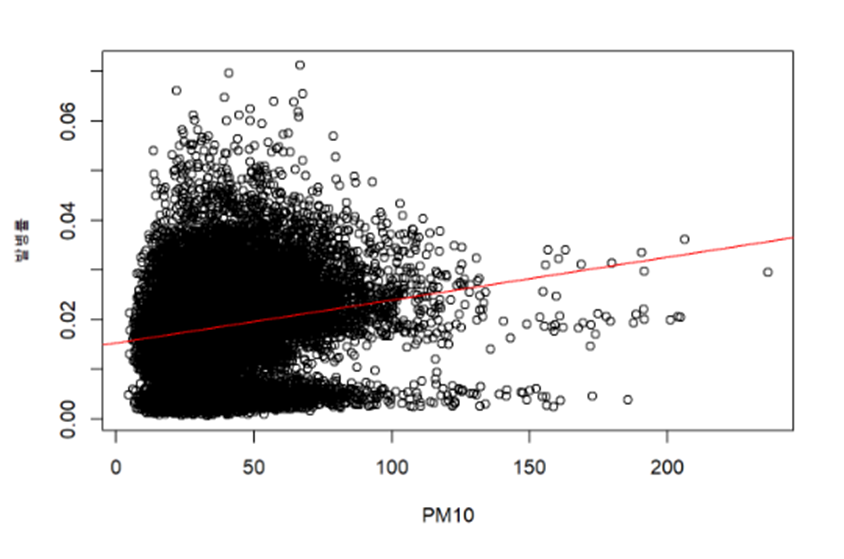


[ 표 2 ] 군집 간 특성 비교

발병률이 높게 나타났던 세종과 발병률이 낮은 타 군집 간의 특성을 비교해본 결과, 세종은 타 군집에 비해 PM2.5, CO, 일최심신적설이 높고, 평균 풍속이 낮은 것으로 나타났다.

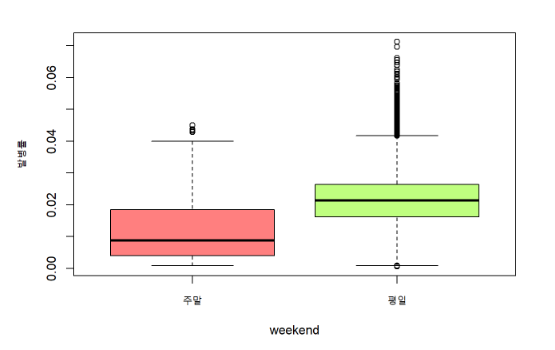
# 회귀분석

## 요인과 발병률 간 상관관계 확인



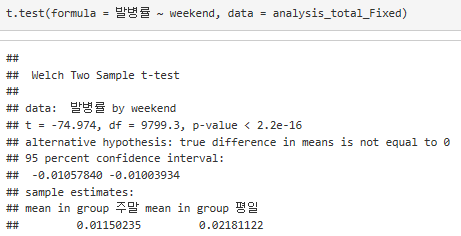
[ 그림 47 ] 미세먼지, 발병률 간 산점도

일별 시계열 자료를 통해 산점도를 그려본 결과, 발병률이 일정 수준 이하로 나타나는 구간이 존재하는 것을 확인할 수 있었다.

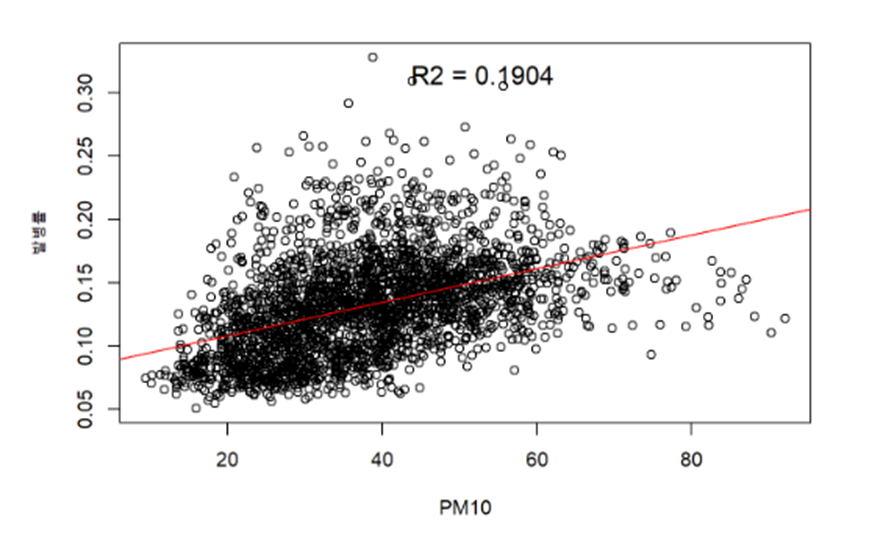


[ 그림 48 ] 평일, 주말 간 발병률 차이

직접 데이터를 확인해본 결과 ,평일과 주말의 발병률 차이로 인해 나타나는 현상임을 확인.이를 T 검정으로 통계적 유의성을 확인.



T-검정 결과, 평일과 주말의 발병률 차이는 통계적으로 유의미한 것으로 나타남.



[ 그림 49 ] 주별 미세먼지, 발병률 간 산점도

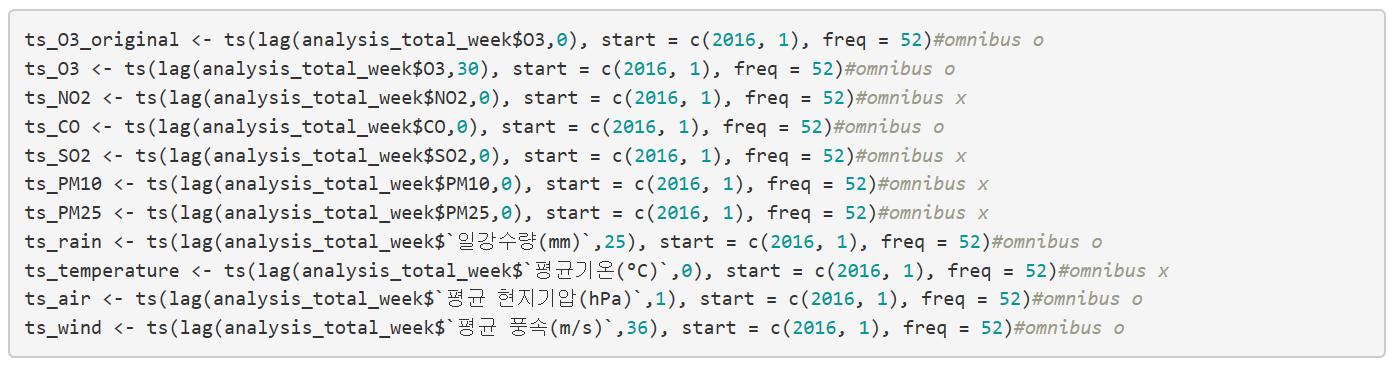
일별이 아닌 주별로 산점도를 다시 그려본 결과, 결정계수의 값이 개선된 것을 확인할 수 있음.

따라서, 일별이 아닌 주별 자료를 사용하여 회귀분석을 진행하기로 결정함.

## 지연효과 적용

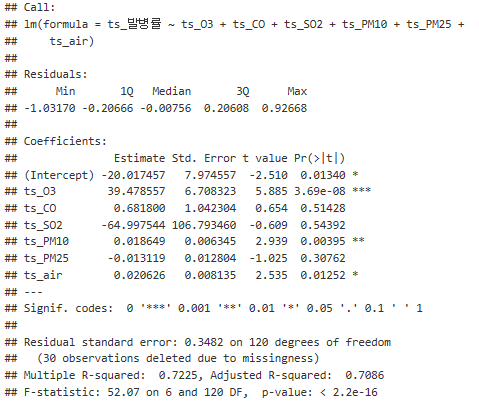
## C:\Users\acorn\Downloads\다운로드.pngC:\Users\acorn\Downloads\다운로드 (1).png

[ 그림 50 ] 지연효과 적용 후 발병률, O3 패턴

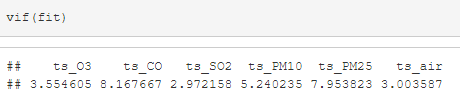


O3와 마찬가지로 일강수량, 평균 풍속 등 지연효과를 적용했을 때, 설명력이 높아진 것을 알 수 있었다. 이를 통해 최적의 lag 값을 적용해서 회귀분석을 진행했음.

7.3 분석 결론



[ 그림 51 ] 회귀분석 결과



발병률과 상관성이 높은 요인들을 추출해서 회귀분석을 진행했다.

NO2의 경우 O3가 NO2와 같은 질소산화물 등이 분해 되어 생성되는 2차 오염물질 임을 고려했을 때, O3와 NO2간의 다중 공선성이 높을 것으로 예상 되었고 회귀분석을 진행했을 때에도 VIF값이 높게 나왔기 때문에 NO2를 제외하고 회귀분석을 진행하였다.

그 결과, O3가 발병률에 미치는 영향력이 통계적으로 가장 유의미하게 나타났고, 다음으로는 평균현지기압, PM10과 평균현지기압 순으로 나타났다.

# 결론 및 제언

## 결론

호흡기질환이 미세먼지에 영향을 받을까?

2015~2018년 뉴스 데이터를 분석해 보았을 때, 호흡기 질환 발생의 원인으로 미세먼지가 가장 많이 언급되고 있는 것을 확인할 수 있다. 빅데이터를 통해 분석한 결과를 보았을 때도, 호흡기 질환 발병률에 영향을 미치는 주요인으로는 O3가 가장 유의미하게 나타났고 그 다음으로 PM10, 평균현지기압이 유의미한 영향력을 주는 것으로 나타났다.

미세먼지와 O3의 경우 이미 선행 연구에서 호흡기 질환과의 연관성을 알아볼 수 있었지만 평균 현지 기압이 호흡기 질환에 미치는 영향력에 대해서는 관련된 선행 연구를 찾을 수 없었다. 이에 평균 현지 기압이 호흡기 질환 발병률에 미치는 영향에 관한 추가적인 연구가 필요할 것으로 보인다.

## 제언

호흡기 질환 발병률을 줄이기 위해서는 분석 결과 호흡기 질환 발병률에 가장 높은 영향력을 미치는 것으로 나타난 O3와 PM10의 발생을 줄일 필요가 있다.

O3 지표 오존은 가정, 자동차, 사업장 등에서 대기 중으로 직접 배출되는 오염물질이 아니라, 이들에서 배출된 질소산화물(NOx = NO+NO2 : 연소과정에서 90∼95% 이상이 NO로 배출됨)과 휘발성유기화합물이 대기 중에 함께 존재한 상태, 또는 질소산화물만이 존재하는 상태에서 햇빛에 의한 광화학반응으로 생성되는 2차적 오염물질이다. 그렇기 때문에 NO2와 휘발성유기화합물을 감소시키는 것이 오존의 감소에 도움을 줄 것으로 판단된다.

NO2, 휘발성 유기화합물의 주요 배출원으로는 자동차와 파워 플랜트와 같은 고온 연소공정과 화학물질 제조공정 등 인위적 현상과, 토양중의 세균에 의해 생성되는 자연적 현상 등으로 파악되고 있다. NO2의 경우에는 도로변이 많은 서울, 경기 밀집지역에서 수치가 가장 높은 것으로 나타났다. 또한, NO2의 경우 2018년도 이후 증가하는 추세를 보이고 있다.

따라서, NO2, 휘발성 유기화합물 농도 개선을 목표로 자동차, 사업장, 에너지 등 광범위한 분야에 걸친 오염물질 배출량 저감정책을 시행하여 궁극적으로 2차 오염물질인 O3의 농도까지 줄일 수 있도록 하여야 한다.

마지막으로 호흡기 질환의 또 다른 주요 원인으로 꼽히는 미세먼지의 경우 중국의 영향이 크게 작용하는 것으로 예측되는 만큼 한국, 중국 양 국가 간 환경협력 대화 채널 및 환경산업 기술협력 등 국가의 국외 대응 방안에 대해 국가가 적극 적으로 참여할 필요가 있다. 또한 이에 대한 정확한 정보를 확보하고 미세먼지 해결에 대해 선도적으로 대응하여 나아갈 필요가 있다.