[2020 시계열분석 및 실습 프로젝트 5 조 보고서]

한국과 중국의 미세먼지(PM10), 미세먼지(PM2.5) 데이터 분석

경제학부 강종윤,
지구환경과학부 한상우
통계학과 김민국
통계학과 한지우

1. 서론

포털 사이트에서 오늘의 날씨를 검색하면, 기온, 강수량 외에 미세먼지, 초미세먼지 수치를 볼 수 있다. 특히 봄과 같이 황사가 심할 때에는 미세먼지, 초미세먼지 수치를 확인하는 사람이 늘었다. 실제로 '미세미세'와 같이 미세먼지 수치를 알려주는 휴대폰 앱도 인기를 끌었고, 많은 사람이 다운받아 사용하기도 하였다. 미세먼지는 공기 중 고체상태와 액체상태의 입자의 혼합물로 배출되며 화학반응 또는 자연적으로 생성된다. 미세먼지는 직경에 따라 PM10 와 PM2.5 로 구분된다. PM 은 Particulate Matter 의 약자이고 뒤에 붙는 숫자가 먼지의 크기를 의미한다. PM10 은 1000 분의 10mm 보다 작은 먼지이며, PM2.5 는 1000 분의 2.5mm 보다 작은 먼지로, 머리카락 직경(약 60 ㎛)의 1/20~1/30 크기보다 작은 입자이다. 이러한 미세먼지는 천식과 같은 호흡기계 질병을 악화시키고, 폐 기능의 저하를 초래한다. PM2.5 는 입자가 미세하여 코 점막을 통해 걸러지지 않고 흡입 시 폐포까지 직접 침투하여 천식이나 폐질환의 유병률과 조기사망률을 증가시킨다. 이러한 이유로, 미세먼지, 초미세먼지 수치가 높으면 마스크를 사용하고 외출하기를 권장한다. 또한 미세먼지는 시정을 악화시키고, 식물의 잎 표면에 침적되어 신진대사를 방해하며, 건축물이나 유적물 및 동상 등에 퇴적되어 부식을 일으킨다.

국제적으로 미세먼지의 유해성이 밝혀지자 1987 년, 세계보건기구(WHO)는 미세먼지(PM10, PM2.5)에 대한 가이드라인을 제시하였고, 2013 년에는 WHO 산하 국제암연구소(IARC)에서 미세먼지를 1 군 발암물질로 지정하였다. 우리나라의 대기환경기준은 세계보건기구(WHO)와 유럽연합(EU)에 비해 높은 수준이다. 그린피스에 따르면, 한국이 OECD 회원국 중 초미세먼지가 가장 심각하다고 한다. 2019 년 3 월에는, 미세먼지가 심각하여 수도권 등 7 개 시도에서 비상 저감 조치를 실시하였다. 많은 언론도 국내 미세먼지 이슈에 대한 관심이 뜨겁다. 한 기사에 따르면, 국민들 중 90%는 미세먼지 오염이 심각하다고 하며, 52%는 발생 원인은 중국 등 국외에 있다고 말했다고 한다. 이번 시계열 프로젝트에서는 많은 사람들이 관심을 가지는 한국의 미세먼지 수치와 초미세먼지 수치를 데이터로 얻은 후 이에 대한 시계열 분석을 실시한다. 또한 중국의 미세먼지 데이터를 이용하여 한국과 유사한 점이 있는지 비교 분석을 실시한다.

시간에 따라 변하는 미세먼지 수치를 분석하기 위해 시계열모형을 사용해야 한다고 생각하였다. 본 프로젝트에서는 시계열 분석 기법 중 시계열(seasonality)이 고려된 ARIMA(autoregressive integrated moving average) 모형인 승법계절 ARIMA 모형(ARIMA(p,d,q)[P,D,Q]s)을 이용하여 분석을 진행하였다. 주어진 데이터를 plotting 하였고, 추이를 보기 위해 smoothing 기법을 이용하였다. 이후 일별 데이터를 월별 데이터로 변환하여 약 70 개의 시계열 데이터를 얻었다. 경향성을 기준으로 적절한 변환도 결정하였으며, adf.test()를 통해 정상성을 확인하였다. 또한, lag plot, decomposition 등을 활용하여 계절차분을 결정하였다. 이후 auto.arima() 방법을 통해 찾은 모형과 AIC, BIC 를 최소로 하는 모형을 찾았다. 3 개의 후보 모형 중 잔차의 자기상관성, 정규성, 모수의 유의성, 모수의 수를 기준으로 최종 모형을 선택하였다. 마지막으로 prediction 을 통해 실제 값을 잘 예측하는지 확인하였다. 이후 결과를 해석하였고, 정부의 정책과 비교하여 정책이 효과가 있었는지에 대해서도 논하였다.

2. 데이터 소개

본 연구에 사용된 한국과 중국의 미세먼지 데이터는 aqicn.org 에서 다운받은 자료이다. 이 사이트는 세계의 실시간 대기질 지수 데이터를 제공하고 있다. Real-time air quality index 인 AQI 정보가 2013 년 12 월 31 일부터 2020 년 5 월 20 일까지 일별로 정리되어 있는데 이를 다운받아 사용하였다. 두 가지 미세먼지 종류인 미세먼지(PM10), 미세먼지(PM2.5) 수치 데이터를 얻었으며, 한국의 경우 서울, 대전, 부산의 미세먼지 수치의 평균, 중국은 베이징, 천진, 상하이, 청도, 심양, 대련의 미세먼지 수치의 평균이다.

데이터의 형태는 [그림 1]과 같다. 그림에서 볼 수 있듯이, 데이터마다 결측치가 존재하고, 특히 한국 미세먼지(PM2.5) 데이터는 2014 년 8 월 2 일 이전까지 모두 결측치이기 때문에 결측치에 대한 처리가 필요하였다. Data 시작 일시는 연속 7 일 이상 데이터의 값이 존재하는 첫 시점으로 결정하였으며, 이후의 Data 에 포함된 NA 값들은 선형적으로 채웠다. 예를 들어, 10 NA NA 7 이라는 데이터가 있다면, 10 9 8 7 로 바꿨다.

	DATE	BEIJING	TIANJIN	SHANGHA	QINGDAO	SHENYAN	DALIAN	CHINA
1	2013-12-31	85	136	121	179	72	97	115
2	2014-01-01	136	216	92	170	94	76	130.6667
3	2014-01-02	79	102	105	189	65	92	105.3333
4	2014-01-03	109	74	94	120	131	54	97
5	2014-01-04	74	114	53	154	101	88	97.33333
6	2014-01-05	91	114	36	367	122	103	138.8333
7	2014-01-06	69	113	48	121	71	111	88.83333
8	2014-01-07	17	55	39	103	48	53	52.5
9	2014-01-08	42	74	52	90	88	54	66.66667
10	2014-01-09	77	168	62	138	106	83	105.6667
11	2014-01-10	120	391	51	175	113	112	160.3333
12	2014-01-11	36	70	66	164	60	75	78.5
13	2014-01-12	97	109	46	93	125	65	89.16667
14	2014-01-13	94	154	25	202	150	74	116.5
15	2014-01-14	162	332	34	320	121	166	189.1667
16	2014-01-15	340	221	69	147	158	160	182.5
17	2014-01-16	88	126	79	242	123	117	129.1667
18	2014-01-17	57	222	112	165	84	84	120.6667
19	2014-01-18	125	239	89	149	116	98	136
20	2014-01-19	30	73	149	264	81	99	116
21	2014-01-20	36	47	73	75	121	41	65.5
22	2014-01-21	102	124	57	91	121	58	92.16667
23	2014-01-22	173	142	41	95	84	78	102.1667

	DATE	BEIJING	TIANJIN	SHANGHA	QINGDAO	SHENYAN	DALIAN	CHINA
1	2013-12-31	NA	NA	NA	NA	NA	NA	NA
2	2014-01-01	125	170	188	182	106	141	152
3	2014-01-02	218	200	170	185	145	124	173.6667
4	2014-01-03	127	157	191	209	134	134	158.6667
5	2014-01-04	213	143	176	136	216	95	163.1667
6	2014-01-05	168	180	116	226	188	157	172.5
7	2014-01-06	204	188	83	346	214	186	203.5
8	2014-01-07	178	182	120	157	141	189	161.1667
9	2014-01-08	65	109	106	144	82	88	99
10	2014-01-09	92	135	97	157	168	95	124
11	2014-01-10	160	221	120	205	179	152	172.8333
12	2014-01-11	218	389	127	246	188	189	226.1667
13	2014-01-12	76	142	152	125	118	115	121.3333
14	2014-01-13	190	178	104	121	212	126	155.1667
15	2014-01-14	190	204	76	181	247	160	176.3333
16	2014-01-15	249	356	94	302	190	250	240.1667
17	2014-01-16	466	320	156	234	265	277	286.3333
18	2014-01-17	228	200	169	282	213	200	215.3333
19	2014-01-18	134	282	207	259	170	147	199.8333
20	2014-01-19	185	287	179	235	202	168	209.3333
21	2014-01-20	42	81	226	233	177	167	154.3333
22	2014-01-21	95	107	122	153	213	85	129.1667
23	2014-01-22	211	188	132	167	209	149	176
24	2014-01-23	339	201	114	174	166	163	192.8333

	DATE	SEOUL	DAEJEON	BUSAN	KOREA
1	2013-12-31	NA	NA	NA	NA
2	2014-01-01	85	87	88	86.66667
3	2014-01-02	53	65	79	65.66667
4	2014-01-03	57	71	65	64.33333
5	2014-01-04	49	53	63	55
6	2014-01-05	43	57	36	45.33333
7	2014-01-06	61	63	66	63.33333
8	2014-01-07	74	66	74	71.33333
9	2014-01-08	65	62	62	63
10	2014-01-09	21	21	44	28.66667
11	2014-01-10	33	32	49	38
12	2014-01-11	49	56	68	57.66667
13	2014-01-12	57	68	81	68.66667
14	2014-01-13	30	32	62	41.33333
15	2014-01-14	42	50	46	46
16	2014-01-15	56	61	55	57.33333
17	2014-01-16	80	73	63	72
18	2014-01-17	94	93	95	94
19	2014-01-18	57	61	77	65
20	2014-01-19	50	41	46	45.66667
21	2014-01-20	60	62	72	64.66667
22	2014-01-21	78	53	62	64.33333
23	2014-01-22	88	60	78	75.33333
24	2014-01-23	57	56	56	56.33333

	DATE	SEOUL	DAEJEON	BUSAN	KOREA
1	2013-12-31	NA	NA	NA	NA
2	2014-01-01	NA	NA	NA	NA
3	2014-01-02	NA	NA	NA	NA
4	2014-01-03	NA	NA	NA	NA
5	2014-01-04	NA	NA	NA	NA
6	2014-01-05	NA	NA	NA	NA
7	2014-01-06	NA	NA	NA	NA
8	2014-01-07	NA	NA	NA	NA
9	2014-01-08	NA	NA	NA	NA
10	2014-01-09	NA	NA	NA	NA
11	2014-01-10	NA	NA	NA	NA
12	2014-01-11	NA	NA	NA	NA
13	2014-01-12	NA	NA	NA	NA
14	2014-01-13	NA	NA	NA	NA
15	2014-01-14	NA	NA	NA	NA
16	2014-01-15	NA	NA	NA	NA
17	2014-01-16	NA	NA	NA	NA
18	2014-01-17	NA	NA	NA	NA
19	2014-01-18	NA	NA	NA	NA
20	2014-01-19	NA	NA	NA	NA
21	2014-01-20	NA	NA	NA	NA
22	2014-01-21	NA	NA	NA	NA
23	2014-01-22	NA	NA	NA	NA
24	2014-01-23	NA	NA	NA	NA

[그림 1] 중국과 한국의 미세먼지(PM10), 미세먼지(PM2.5) 수치. 위의 두 개의 표는 각각 중국의 미세먼지(PM10), 미세먼지(PM2.5) 수치로, 베이징, 천진, 상하이, 청도, 심양, 중국 지역의 수치가 포함되었다. 아래 두 개의 표는 각각 한국의 미세먼지(PM10), 미세먼지(PM2.5) 수치로 서울, 대전, 부산, 한국 지역의 수치가 포함되었다. 각 데이터는 NA 값을 포함하고 있으며, 특히 한국 미세먼지(PM2.5) 데이터의 경우 초반 데이터에 NA 값이 많다.

데이터의 소개는 [표 2]와 같다. 위에서 설명한 방식처럼 결측치를 처리한 이후, 일별 데이터의 시작 날짜와 데이터 수, 결측치 수를 정리한 결과이다.

	시작 날짜	데이터 수	결측치 수
중국 PM10	2013-12-31	2333	47
중국 PM2.5	2014-01-01	2332	12
한국 PM10	2014-01-01	2332	71
한국 PM2.5	2014-08-02	2119	71

[표 1] 결측치 처리 이후 일별 데이터의 소개. 각 데이터별로 데이터의 시작 날짜와 시작 날짜 이후, 데이터의 개수와 결측치 개수를 포함하고 있다.

이후 일별 데이터를 월별 데이터로 바꾸어, 최종 분석 데이터로 정제하였다. 각 월별 평균을 계산하였으며, 첫 달과 마지막 달의 경우 모든 날의 데이터가 있지 않아서 이를 제외하고, 최종 데이터를 결정하였다. 중국 미세먼지(PM10)의 경우 2014 년 1 월부터 총 76 개의 데이터가 남았으며, 중국 미세먼지(PM2.5)와 한국 미세먼지(PM10)의 경우 2014 년 2 월부터 총 75 개의 데이터가 남았다. 마지막으로 한국 미세먼지(PM 2.5)의 경우 68 개의 데이터가 남았다.

3. 연구 방법

3.1. Data 확인 및 Transformation 결정

3.1.1. Data plot 을 통한 경향성 확인

경향성을 확인할 때에는 일별 데이터를 이용하였다. *ggplot2* 패키지를 활용하여 추이를 확인하였는데, 조금 더 경향성을 잘 확인하기 위해 추가적으로 geom_smooth() 함수를 이용하여 GAM(Generalized Additive Model)으로 smoothing 하였다. GAM 은 선형모형에 비선형 함수의 additivity 를 허용하는 회귀 모형이며, REML estimation 을 이용한 모형이다. 또한, 옵션을 변경하여 Loess 방식으로도 smoothing 하였는데, 국소회귀라고 불리는 이 방법은 근처 값들에 가중치를 적용하여 weighted least squares estimation 으로 회귀 계수를 찾는다.

이렇게 그린 결과를 전체, 연도별, 월별로 그려보며, 미세먼지(PM10)과 미세먼지(PM2.5)의 경향성을 확인하고 비교하였다. 또한, 한국과 중국의 경향성 역시 비교하였다.

3.1.2. Transformation 을 통한 분산 안정화

이후 분석을 월별 데이터로 진행하였고, 이 때 월별 데이터의 경향성도 확인해보았다. 이 때, 분산 안정화 작업이 필요하다고 판단되는 데이터에 분산 안정화를 위해 log transformation 을 진행하였다. 미세먼지 수치가 높은 시기에는 분산이 비교적 크고, 미세먼지 수치가 낮은 시기에는 분산이 비교적 작기 때문에 log transformation 을 이용하면, 분산 안정화 효과를 볼 수 있을 것이라 판단하였다.

3.2. 모형 결정

3.2.1. 차분 및 계절차분의 결정

먼저, Augmented Dickey-Fuller Test(adf.test())를 이용하여 단위근(unit root)을 가지는지 판단하였다. 이후, 3.1.1.에서 그린 그래프와 decomposition 결과를 이용하여 trend 가 있는지 확인하고, 이를 토대로 차분 여부를 결정하였다. 또한, 미세먼지의 경우 계절성을 가짐을 짐작할 수 있었고, 실제 데이터를 통해서 확인해보고, 계절 차분 여부 및 주기를 결정하였다.

이를 위해 Lag plot 을 그려서 계절차분을 결정하였다. Lag plot 의 경우 원래 시계열 데이터와 일정 기간만큼 시기를 shift 한 데이터를 시각적으로 표현하는 그래프이며, 상관관계가 높은 주기를 확인할 수 있다. 이 때 periodogram, decomposition plot 의 seasonal 파트와 SACF, SPACF plot 도 이용하여 주기를 최종 결정하였다. 이 때 periodogram 은 *sarima* 패키지의 함수인 periodogram() 함수를 이용하였다. 즉, 이러한 과정을 통해 각 데이터별로 ARIMA 모형의 d, D, s 를 결정하였다.

3.2.2. ARIMA(p,d,q)(P,D,Q)s 모형 적합

ARIMA(p,d,q)(P,D,Q)s 모형을 이용하여 데이터들을 적합하였다. 이 때, 3.2.1.에서 d, D, s 를 결정하였기 때문에 나머지 p, q, P, Q 를 결정하고, 이를 이용한 ARIMA(p,d,q)(P,D,Q)s 모형을 적합하면 충분하였다. 본 프로젝트에서는 총 3 가지 방법으로 p, q, P, Q 를 결정하였다. 첫 번째로는 *forecast* 패키지에 포함된 auto.arima() 함수를 이용하여 결정하였다. 하지만 auto.arima() 함수를 통해 구해진 모형이 AIC 나 BIC 를 최소로 하는 모형은 아니기 때문에, 다른 함수를 만들어서 low AIC, low BIC 모형도 찾았다. p, q, P, Q 의 값을 0 부터 앞에서 auto.arima() 함수로 결정된 p, q, P, Q 의 값에서 1 을 더한 값까지 각각 모형 적합을 한 후 가장 AIC, BIC 가 낮은 모형을 선택하였다. 이 때, [그림 2]와 같이 searcher() 함수를 만들어서 이용하였다. 데이터와, 확인해볼 모수의 리스트, period 인 s 값을 넣어주면, low AIC, low BIC 모형과 p, q, P, Q 값을 return 하는 함수이다.

```
searcher <- function(dat, params, period){
  search <- function(p,q,P,Q){
    model <- Arima(dat, order=c(p,0,q), seasonal=list(order=c(P,1,Q), period=period))
    return(model)
  }
  safe_search <- safely(search)
  result <- params %>% pmap(safe_search)
  result <- result %>% transpose %>% .$result
  x <- which(result %>% map_dbl(length)==0)
  if(length(x)!=0) {result <- result[-x]}
  low_aic <- result %>% map_dbl("aic") %>% which.min()
  low_bic <- result %>% map_dbl("bic") %>% which.min()
  return(list(low_aic = result[[low_aic]],low_aic_par = (result[[low_aic]]$arma) ,
              low_bic=result[[low_bic]], low_bic_par = (result[[low_bic]]$arma)))
}
```

[그림 2] low AIC, low BIC 모형을 찾아주는 함수. Data 와 p, q. P, Q 를 넣어주면, AIC, BIC 가 최소인 모형과 그 때의 p, q, P, Q 값을 return 하는 함수이다.

3 가지 방법으로 결정된 모형들 중 하나를 선택하는 과정은 모형 진단 과정이 가장 잘 맞는 모형을 최종적으로 선택하였다.

3.2.3. 모형 진단을 통한 최종 모형 결정

3.2.2.에서 선택된 모형들에 대해 잔차 분석을 진행하였다. 잔차 분석을 통해, 모형의 가정을 잘 만족하는지 확인하였다. Ljung-Box 의 Q 통계량을 구하여, 잔차들이 자기 상관성을 가지는지 확인하였으며 QQ plot 을 그려 잔차가 정규분포를 따르는지 확인하였다. 또한 모형의 모수들이 유의한지 확인하였다. 실제 분석을 진행할 때에는 analysis 라는 함수를 만들어서 사용하였고, 이는 첨부된 코드에 포함되어 있다. 3 개의 모형(auto.arima(), low AIC, low BIC) 중 잔차들이 자기상관성을 띄지 않고, 잔차의 분포가 정규분포에 가까우며, 모수가 유의한 모형을 선택하였다. 이 때, 결과가 비슷한 경우 모수의 수가 적은 모형으로 최종 모형 결정하였다.

3.3. 예측

3.3.1. Forecast 함수를 이용한 예측

위에서 선택한 ARIMA(p,d,q)(P,D,Q)s 모형이 어느 정도 잘 예측하는지 확인하기 위하여, *sarima* 패키지의 함수인 forecast() 함수를 이용하였다. 이 함수를 이용하여 [그림 3]과 같이 forecast_process() 함수를 만들고, prediction 과 one-ahead prediction 을 진행하였다. 2 년 이전의 데이터로 최근 2 년을 예측하였다.

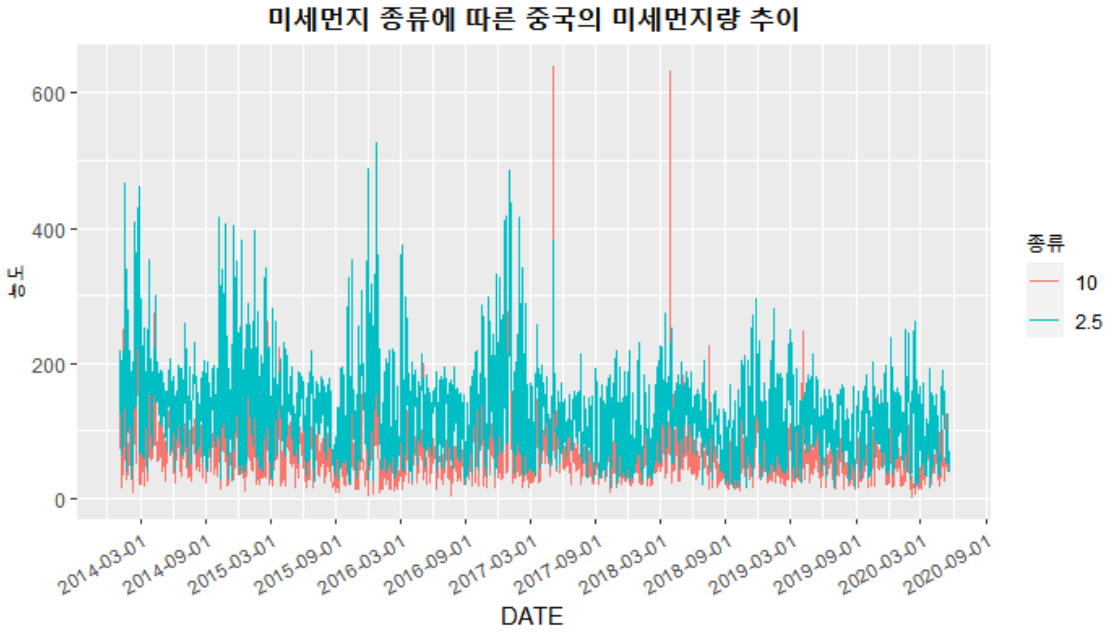
```
forecast_process<-function(data, order, h=12, main=''){
  title=c("Prediction of",'One-ahead Prediction of')
  order.arima = order[1:3]
  order.s = list(order=order[5:7], period = order[4])
  fit = Arima(data[1:(length(data)-h)], order = order.arima, seasonal = order.s)
  print(fit)
  fit2 = Arima(data[(length(data)-h-11):length(data)], model=fit)
  # h-ahead forecast
  par(mfrow=c(2,1),ask=F)
  f = forecast(fit,h=h)
  plot(f,main=paste(title[1],main),ylim=c((min(data)*0.5),(max(c(data, 100))+5)))
  points((length(data)-h):length(data),data[(length(data)-h):length(data)], type='l')
  grid()
  print(f$mean[(length(data)-h):length(data)])
  # one-ahead
  plot(ymd(paste(c(rep(2014,(length(data)-4-60)),rep(2015:2019,each=12),rep(2020,4)),
                 c((13-(length(data)-4-60)):12,rep(1:12,5),1:4),1, sep='-')),data,
       type='l',main=paste(title[2],main),ylim=c((min(data)*0.5),
                                                 (max(c(data, 100))+5)),xlab='date')
  points(ymd(paste(c(rep(2018,8),rep(2019,12),rep(2020,4)),c(5:12,1:12,1:4),1, sep='-')),
         fitted(fit2)[13:(12+h)], col='blue', type='l',lwd=1.5)
  grid()
}
```

[그림 3] forecast 과정 진행 후 plot 을 출력하는 함수. Data 와 order 와 제목을 넣어주면, prediction 과 one-ahead prediction 의 결과를 plot 으로 그려주는 함수이다.

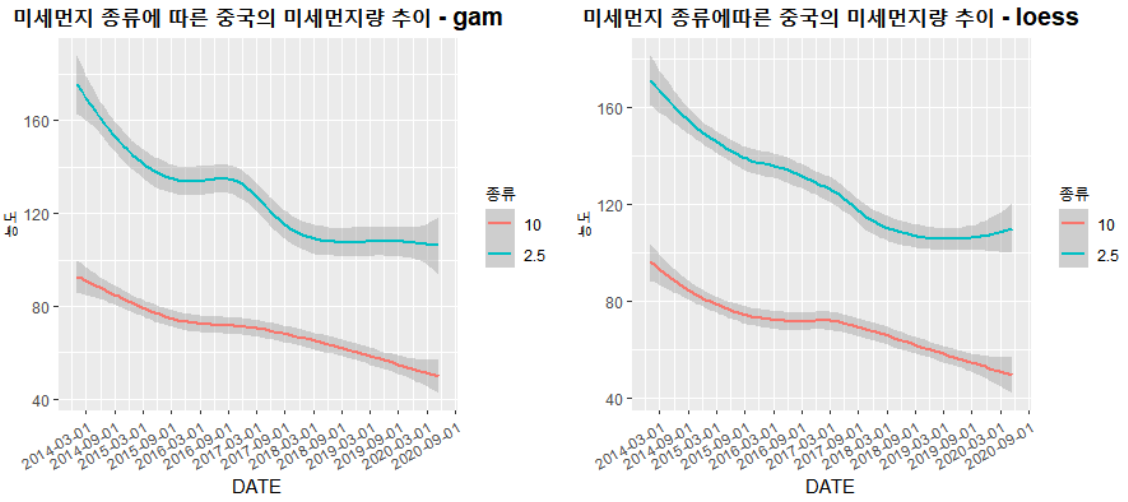
4. Result

4.1. Data 확인 및 Transformation 결정

4.1.1. Data plot 을 통한 경향성 확인

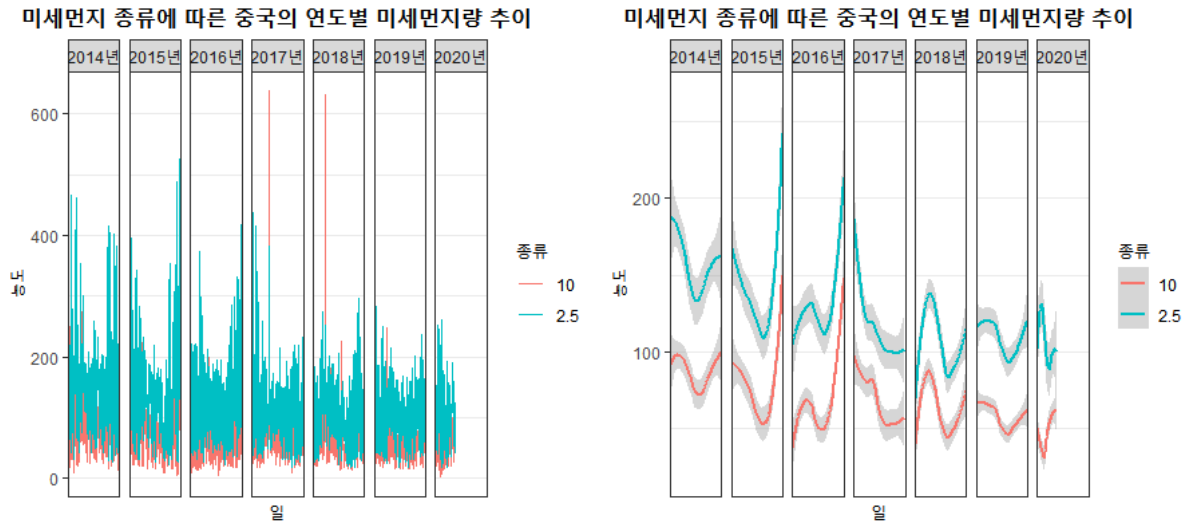


[그림 4] 미세먼지 종류에 따른 중국의 미세먼지량 추이(전체). 일별 데이터에서 중국의 미세먼지(PM10), 미세먼지(PM2.5) 수치를 plot 으로 그렸다.



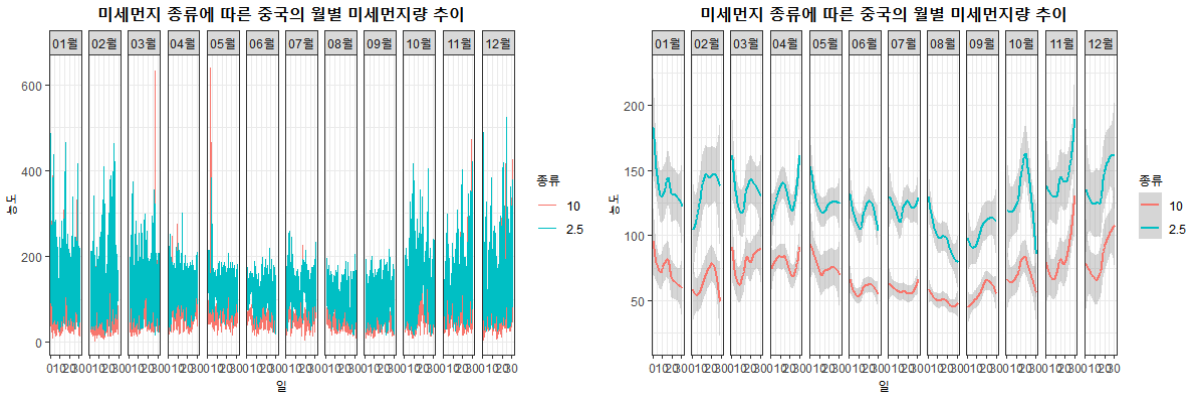
[그림 5] 미세먼지 종류에 따른 중국의 미세먼지량 추이 (gam, loess; 전체). 일별 데이터에서 중국의 미세먼지(PM10), 미세먼지(PM2.5) 추이를 plot 으로 그린 결과이다.

[그림 4], [그림 5]를 보면, 중국의 경우 미세먼지(PM10, PM2.5) 모두 감소하는 경향을 보였다. 봄에서 여름으로 갈수록 감소하는 경향이 있지만, 뚜렷하지는 않았다. 2016 년에 잠시 증가한 경향을 보였다. 더 자세한 경향을 보기 위해 연도별 미세먼지량 추이를 그려보면 다음과 같다.



[그림 6] 미세먼지 종류에 따른 중국의 연도별 미세먼지량 추이. 일별 데이터에서 중국의 미세먼지(PM10), 미세먼지(PM2.5) 수치를 연도별로 나타낸 plot 이 왼쪽 plot 이며, loess 방법으로 smoothing 한 plot 이 오른쪽 plot 이다.

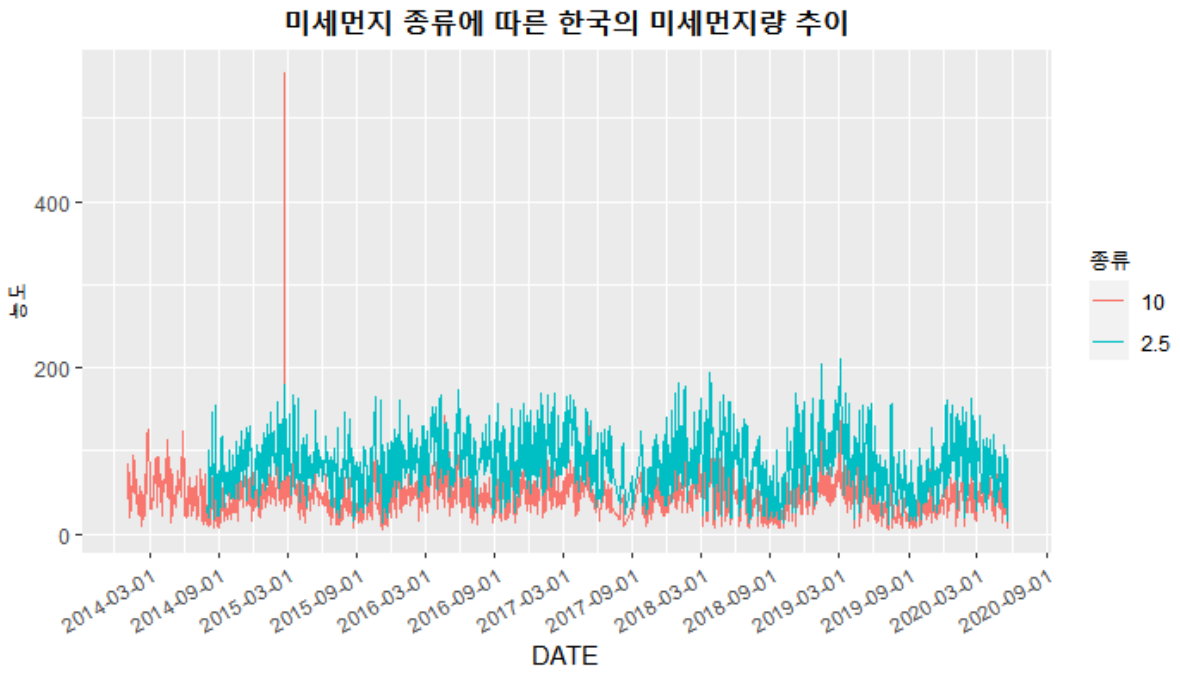
[그림 6]은 중국의 연도별 미세먼지량 추이를 보여주는 그래프들이며, 오른쪽 그래프의 경우 loess 방법으로 smoothing 한 결과이다. 미세먼지(PM10)과 미세먼지(PM2.5)가 비슷한 경향을 가짐을 확인할 수 있었다. 또한, 각 연도별로 경향성을 확인해본 결과, 2014 년과 2015 년에는 감소하다가 증가하는 경향을 2016, 2018, 2019 년에는 증가하다가 감소하는 경향, 2017 년에는 꾸준히 감소하는 경향을 보였다.



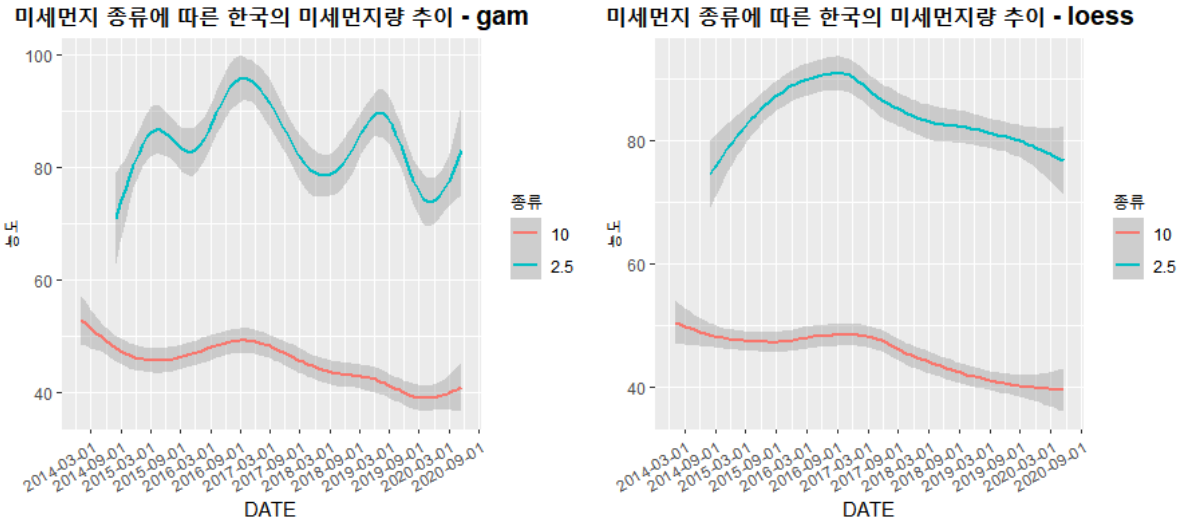
[그림 7] 미세먼지 종류에 따른 중국의 월별 미세먼지량 추이. 일별 데이터에서 중국의 미세먼지(PM10), 미세먼지(PM2.5) 수치를 월별로 나타낸 plot 이 왼쪽 plot 이며, loess 방법으로 smoothing 한 plot 이 오른쪽 plot 이다.

[그림 7]은 중국의 월별 미세먼지량 추이를 보여주는 그래프이다. 연도별 그래프와 유사하게, 미세먼지(PM10), 미세먼지(PM2.5)가 비슷한 경향을 가짐을 확인할 수 있었다. 또한, 각 월별로 경향성을 확인해본 결과, 5 월부터 8 월까지 미세먼지 수치가 꾸준히 감소하였고, 8 월과 9 월에는 다른 달에 비해 미세먼지 농도가 낮은 것으로 확인되었다. 또한, 9 월부터 12 월까지는 다시 미세먼지 수치가 꾸준히 증가하는 추이를 볼 수 있었다. 계절성을 확인할 수 있었다.

이번에는 한국 데이터를 이용하여 경향성을 확인해보는 같은 과정을 반복한 결과이다.

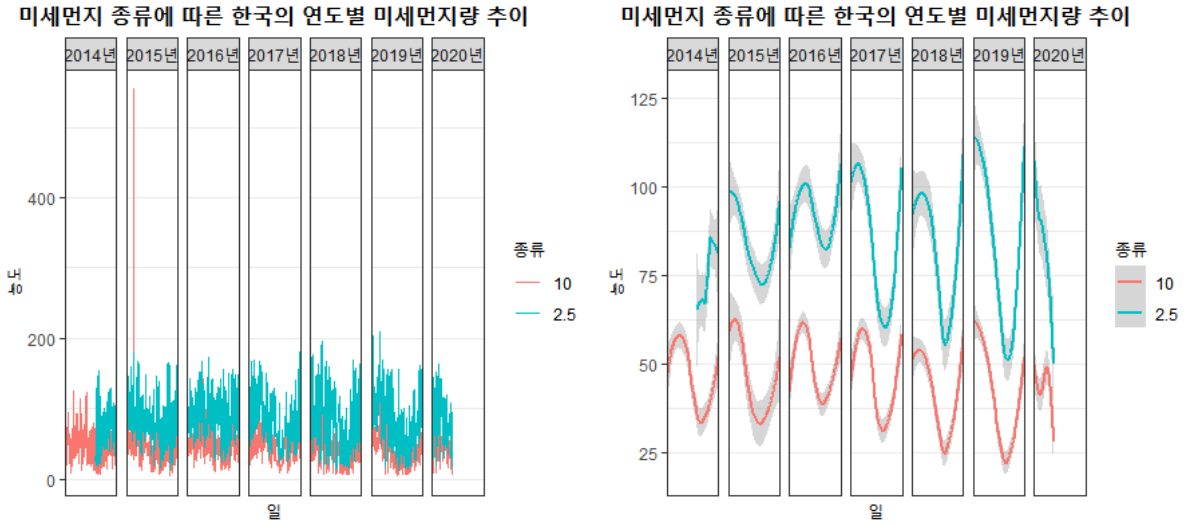


[그림 8] 미세먼지 종류에 따른 한국의 미세먼지량 추이(전체). 일별 데이터에서 한국의 미세먼지(PM10), 미세먼지(PM2.5) 수치를 plot 으로 그렸다.



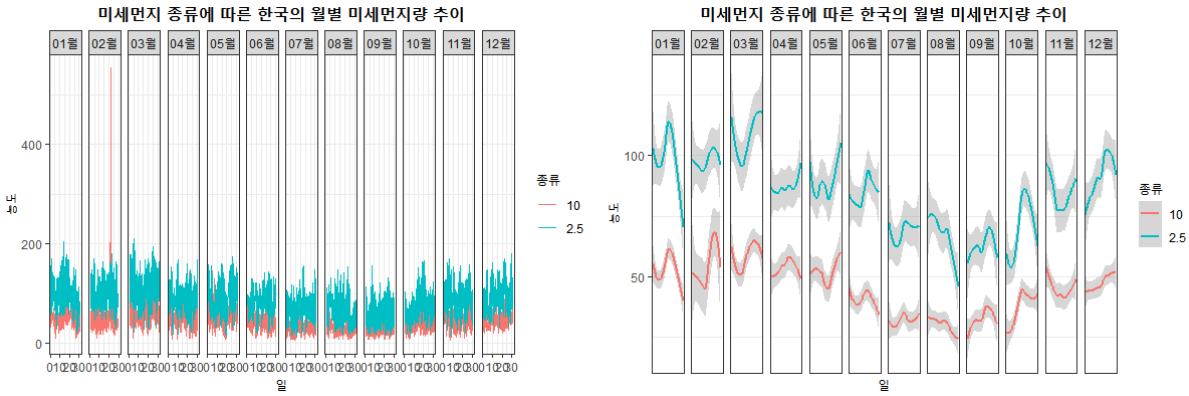
[그림 9] 미세먼지 종류에 따른 한국의 미세먼지량 추이 (gam, loess; 전체). 일별 데이터에서 한국의 미세먼지(PM10), 미세먼지(PM2.5) 추이를 plot 으로 그린 결과이다.

[그림 8], [그림 9]를 보면, 미세먼지(PM10)은 겨울에서 봄, 여름부터 겨울까지는 증가하며, 봄부터 여름까지는 감소하는 경향을 보였다. 미세먼지(PM2.5)는 겨울에서 여름까지 감소하고, 여름부터 겨울까지는 증가하였다. 미세먼지(PM10)의 경우 거의 일정한 수치를 나타낸 반면, 미세먼지(PM2.5)는 2016 년 9 월까지 증가한 후 소폭 감소하였고, 그 이후 일정하게 나타났다.



[그림 10] 미세먼지 종류에 따른 한국의 연도별 미세먼지량 추이. 일별 데이터에서 한국의 미세먼지(PM10), 미세먼지(PM2.5) 수치를 연도별로 나타낸 plot 이 왼쪽 plot 이며, loess 방법으로 smoothing 한 plot 이 오른쪽 plot 이다.

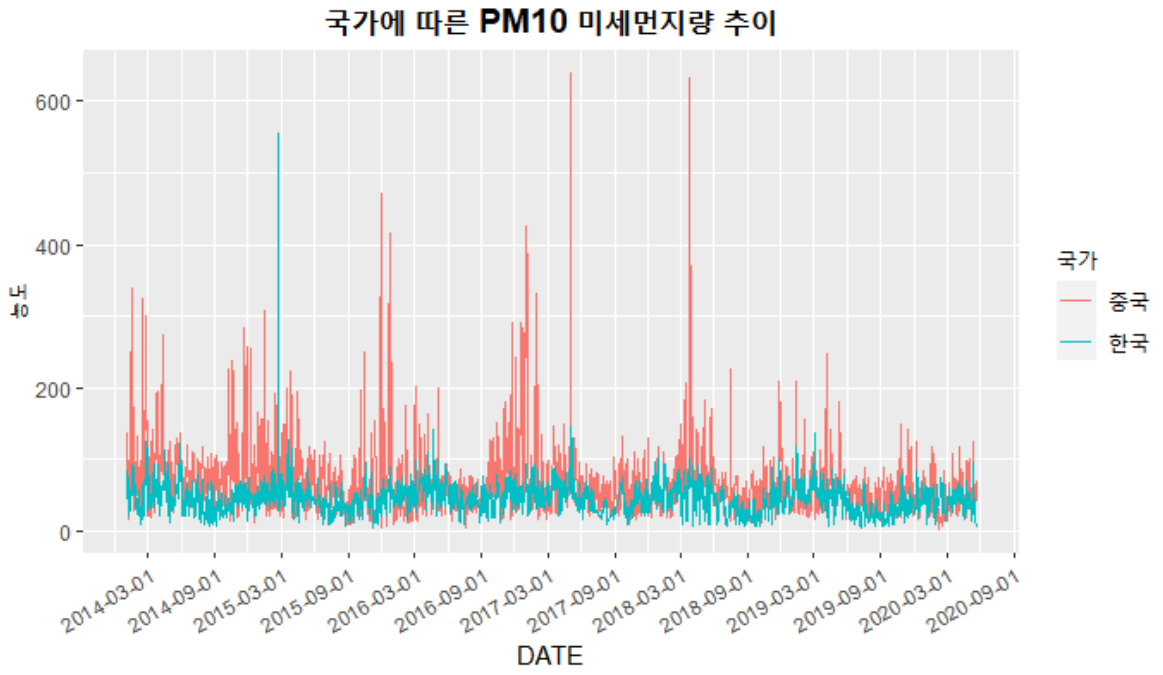
[그림 10]은 한국의 연도별 미세먼지량 추이 그래프들이다. 중국 데이터와 마찬가지로, 미세먼지(PM10)과 미세먼지(PM2.5)는 유사한 경향을 보이고 있다. 또한, 한국 데이터의 경우 모든 연도에서 증가하다가 감소한 후 다시 증가하는 경향을 보이고 있다. 뚜렷한 계절성이 보인다고 할 수 있다.



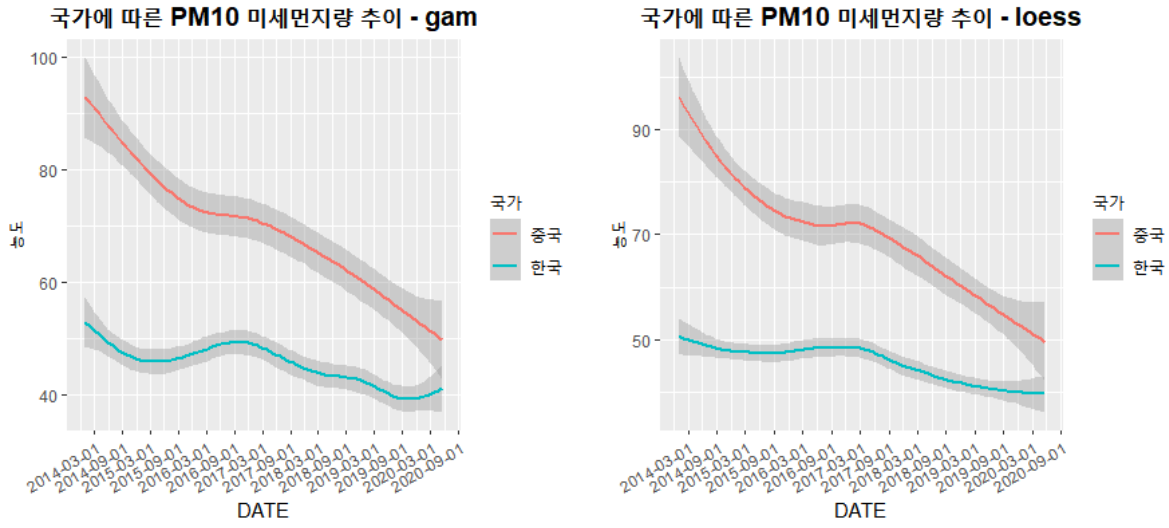
[그림 11] 미세먼지 종류에 따른 한국의 월별 미세먼지량 추이. 일별 데이터에서 한국의 미세먼지(PM10), 미세먼지(PM2.5) 수치를 월별로 나타낸 plot 이 왼쪽 plot 이며, loess 방법으로 smoothing 한 plot 이 오른쪽 plot 이다.

위 [그림 11]은 한국의 월별 미세먼지량 추이 그래프들인데, 역시 미세먼지(PM10)과 미세먼지(PM2.5)는 유사한 경향을 보이고 있었다. 또한, 월별로 경향성이 어떠한지 확인해보면, 7 월부터 9 월까지는 다른 달에 비해 미세먼지 농도가 낮은 것으로 보여졌다. 5 월부터 9 월까지는 꾸준히 감소하는 경향이 나타났으며, 9 월부터 12 월까지는 꾸준히 상승하는 경향을 볼 수 있었다. 계절성을 확인할 수 있었다.

이번에는 중국과 한국의 국가에 따른 미세먼지량 추이를 확인해본다. 먼저, 미세먼지(PM-10) 추이 plot 부터 제시한다.

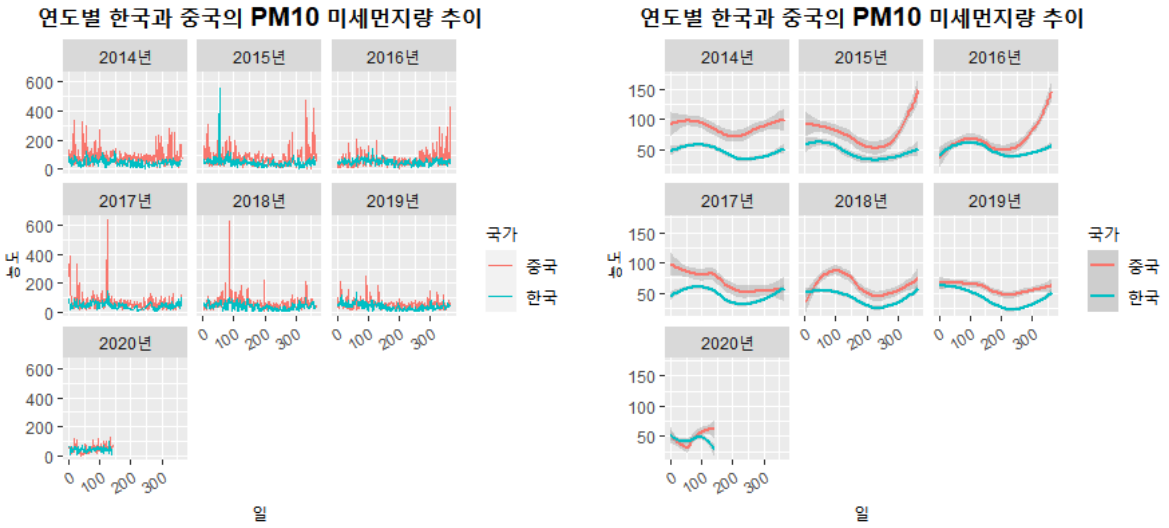


[그림 12] 국가에 따른 PM10 미세먼지량 추이. 중국과 한국의 미세먼지(PM10) 수치를 plot 으로 그렸다.



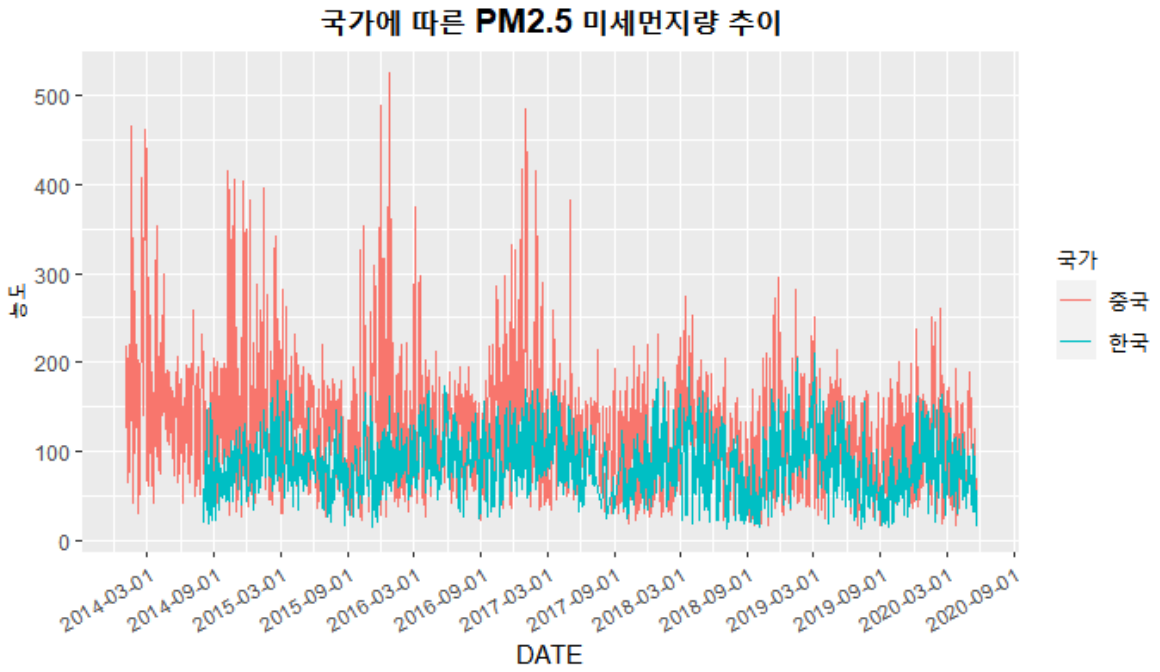
[그림 13] 국가에 따른 PM 10 미세먼지량 추이 (gam, loess). 중국과 한국의 미세먼지(PM10) 추이를 plot 으로 그렸다.

[그림 12], [그림 13]은 중국과 한국의 미세먼지(PM10) 추이를 비교한 그래프이다. 중국의 미세먼지(PM10) 수치는 감소하는 경향을 보이지만, 한국의 미세먼지(PM10) 수치는 거의 일정하거나 조금 감소하는 경향을 보였다. 또한, 중국의 미세먼지(PM10) 수치는 2016 년에 다시 증가하다가 다시 감소하지만, 한국의 미세먼지 수치는 이에 비하면 일정한 편이었다. 하지만 이 그래프로는 둘 다 뚜렷한 계절성을 확인하지는 못했다.



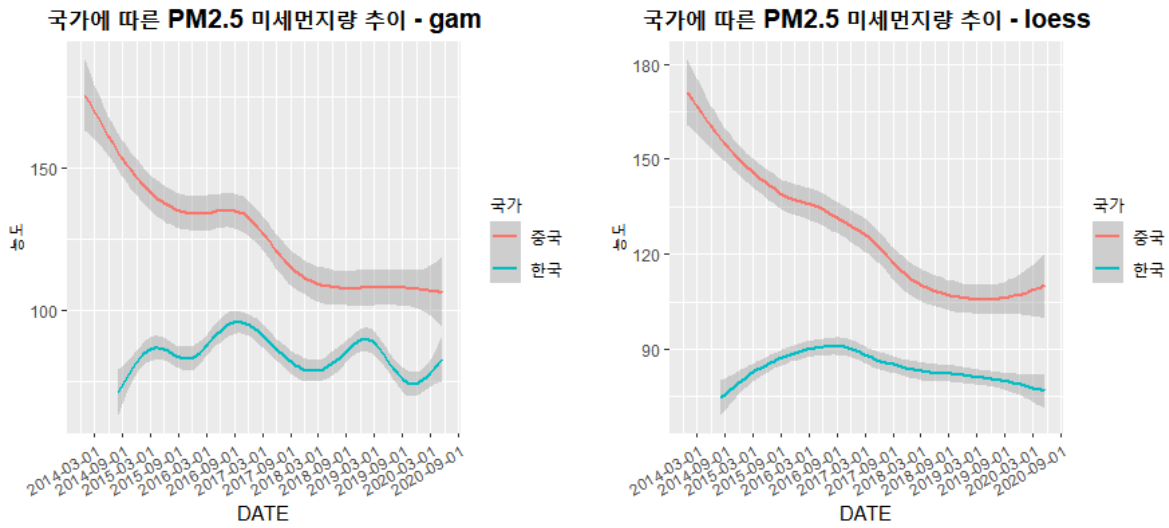
[그림 14] 연도별 한국과 중국의 미세먼지(PM10) 추이. 한국과 중국의 연도별 미세먼지(PM10) 추이를 그래프로 그렸다.

중국과 한국의 미세먼지(PM10) 연도별 경향성은 비슷하게 나타났다. 이번에는 미세먼지(PM2.5) 데이터에 대해 경향성을 확인하면 다음과 같다.



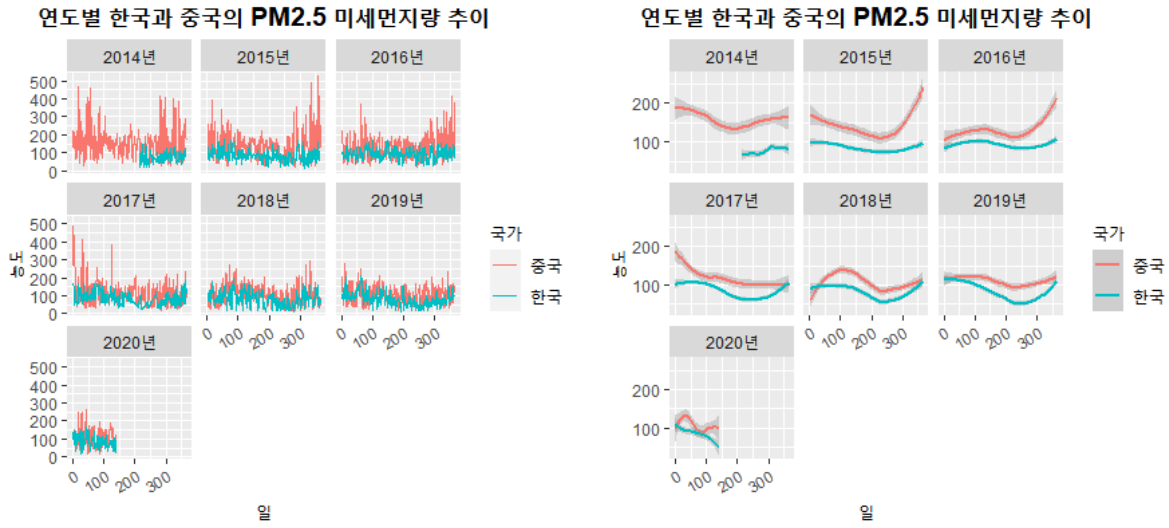
[그림 15] 국가에 따른 PM2.5 미세먼지량 추이. 중국과 한국의 미세먼지(PM2.5) 수치를 plot 으로 그렸다.

위의 [그림 15]에서 plotting 한 미세먼지량 데이터를 gam 과 loess 방법으로 smoothing 한 결과를 [그림 16]에 나타내었다.



[그림 16] 국가에 따른 PM2.5 미세먼지량 추이 (gam, loess). 중국과 한국의 미세먼지(PM2.5) 추이를 plot 으로 그렸다.

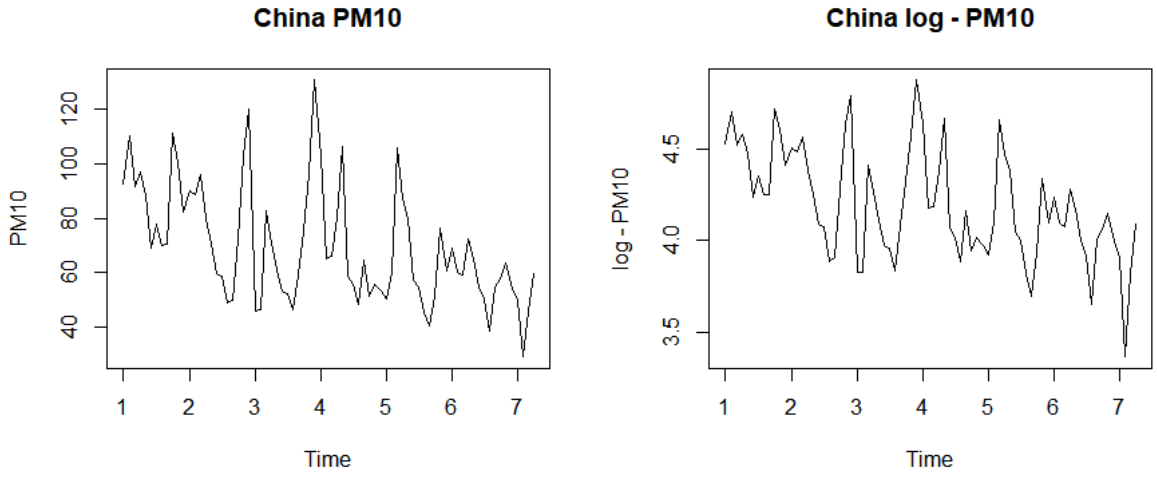
[그림 15], [그림 16]을 보았을 때, 중국의 미세먼지(PM2.5)는 감소하는 경향을 보이지만 한국의 미세먼지(PM2.5)는 일정한 경향을 보인다. 중국의 미세먼지(PM2.5)는 뚜렷한 계절성을 찾기 힘들지만, 한국의 미세먼지(PM2.5)는 봄에서 가을까지는 증가하는 경향이 있었으며, 가을부터 다음 해 봄까지 다시 감소하는 경향이 있었다.



[그림 17] 연도별 한국과 중국의 PM2.5 미세먼지량 추이. 한국과 중국의 연도별 미세먼지(PM2.5) 추이를 그래프로 그렸다.

[그림 17]에서 보면, 앞서 미세먼지(PM10)에서 본 결과와 비슷하게 중국과 한국의 미세먼지(PM2.5) 추이는 거의 유사하게 나타났다. 조금 다른 부분도 있었지만, 어느 정도 경향성이 비슷한 구간들이 있었다.

4.1.2. Transformation 을 통한 분산 안정화



[그림 18] 중국 미세먼지(PM10)의 log transformation. 위 그래프는 미세먼지(PM10) 데이터 plot 이며 아래는 log transformation 한 plot 이다.

왼쪽의 plot 보다 오른쪽의 plot 이 분산이 더 안정적임을 확인할 수 있었다. 분산 안정화를 목적으로 중국 미세먼지(PM10) 데이터에는 log transformation 을 진행하였다.

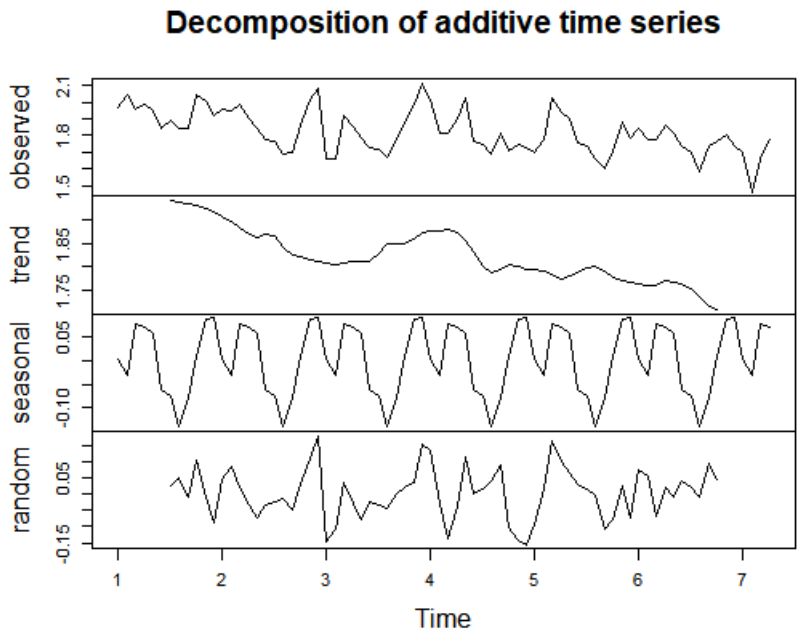
4.2. 모형 결정

4.2.1. 차분 및 계절차분의 결정

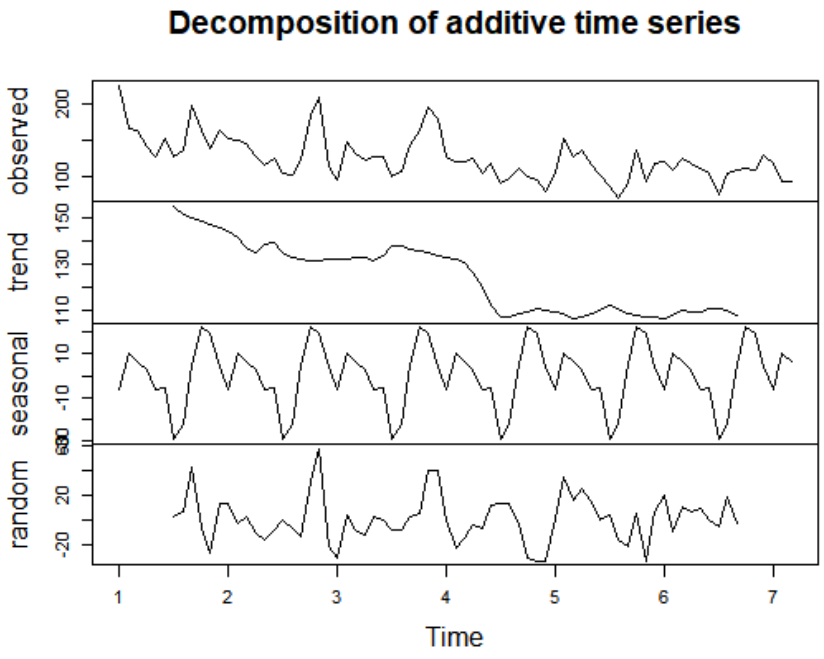
```
        Augmented Dickey-Fuller Test

data:  data
Dickey-Fuller = -4.2175, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

[그림 19] 중국 미세먼지(PM10) 데이터의 adf.test() 결과. 중국 미세먼지(PM10) 데이터에 log transformation 후 adf.test()를 적용한 결과이다. 유의확률이 0.01 로 나타났기 때문에 귀무가설을 기각할 수 있고, alternative hypothesis 인 stationary 가 보장된다.

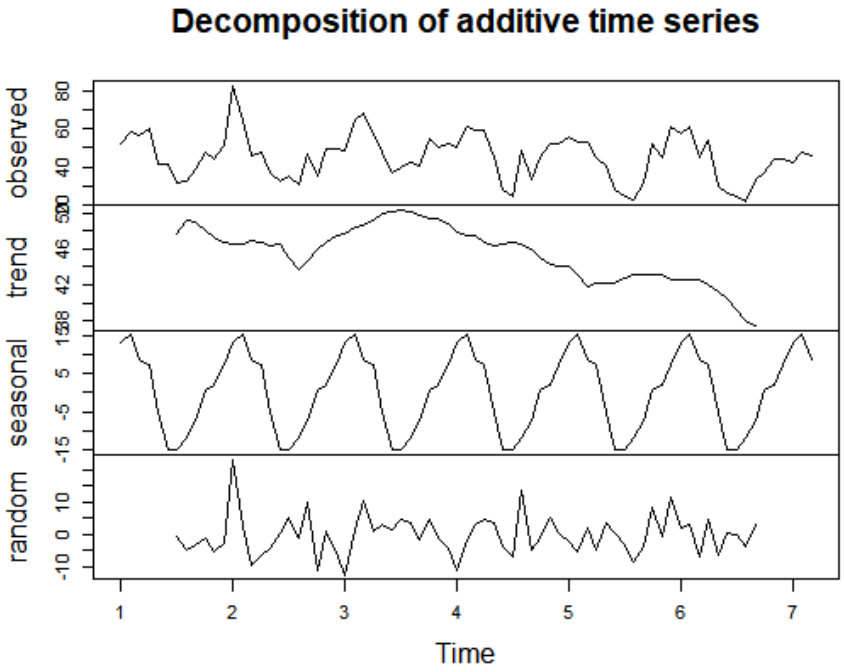
먼저, 정상성 확인을 위해 adf.test()를 해본 결과, 중국의 미세먼지(PM10), 미세먼지(PM2.5), 한국의 미세먼지(PM10), 미세먼지(PM2.5)의 경우 모두 귀무가설을 기각하였기 때문에, 정상성을 만족한다고도 볼 수 있었다. 하지만, Augmented Dickey-Fuller test 의 결과뿐만 아니라 자료를 해석하고 trend 가 있다면 차분을 해야 한다고 생각하였다. 앞서 4.1.1.에서 trend 를 확인할 수 있었지만, 더 정확하게 확인하기 위해 decomposition plot 을 추가로 그렸다.



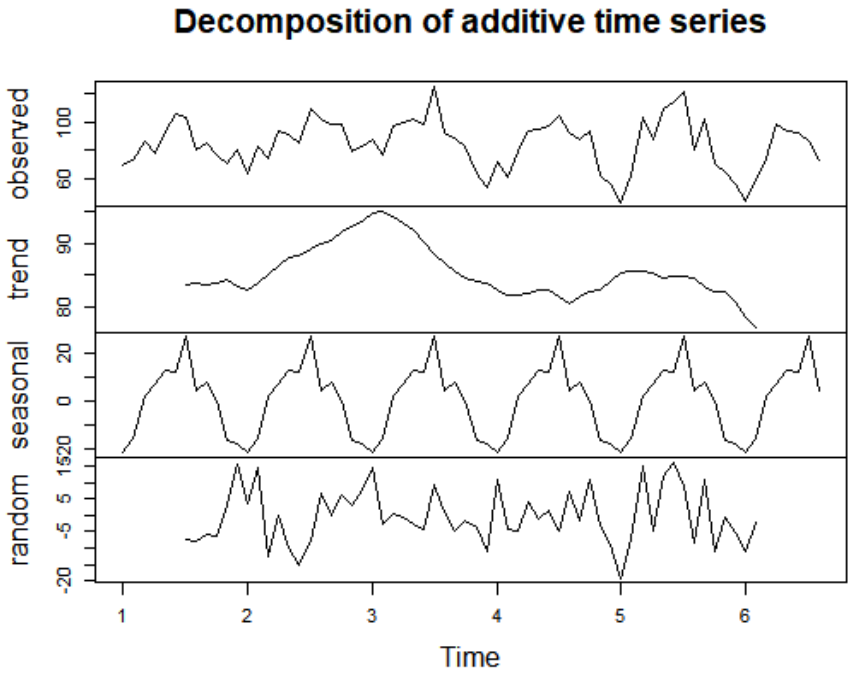
[그림 20] 중국 미세먼지(PM10) 데이터의 decomposition plot. 중국 미세먼지(PM10)에 log transformation 후, decomposition plot 을 그린 결과이다.



[그림 21] 중국 미세먼지(PM2.5) 데이터의 decomposition plot. 중국 미세먼지(PM2.5) 데이터의 decomposition plot 을 그린 결과이다.



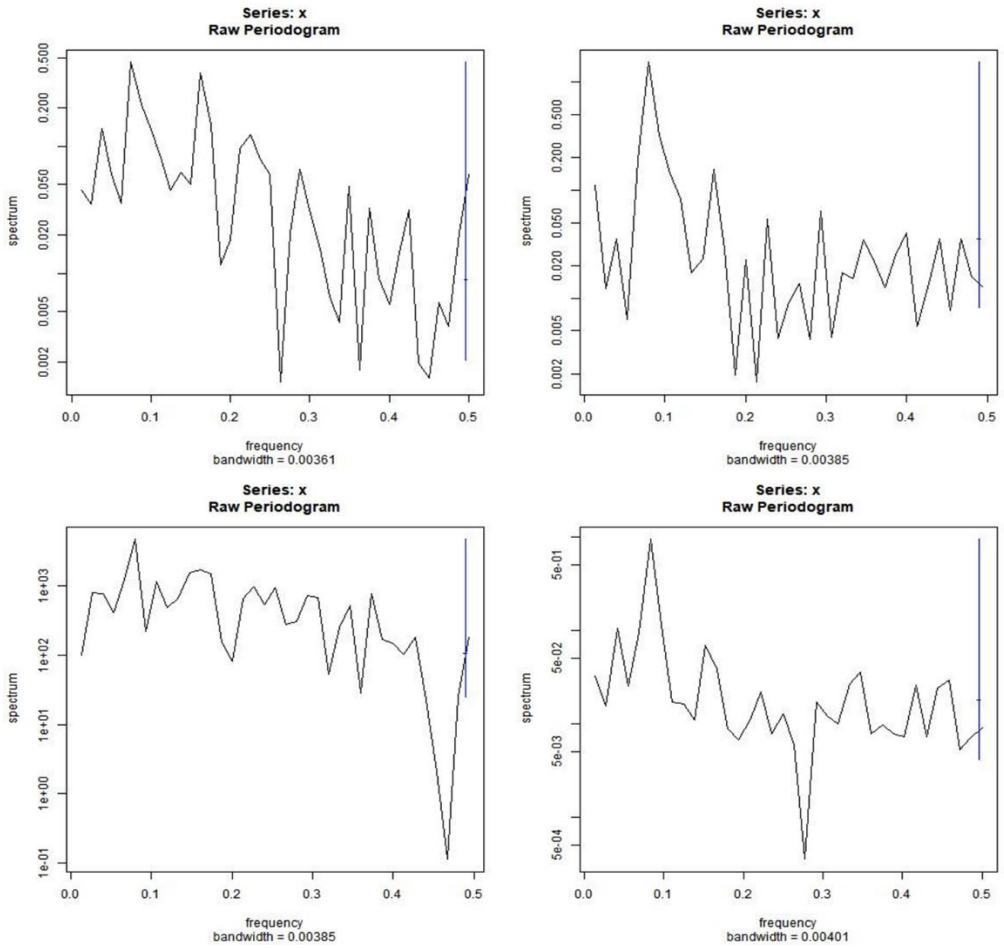
[그림 22] 한국 미세먼지(PM10) 데이터의 decomposition plot. 한국 미세먼지(PM10) 데이터의 decomposition plot 을 그린 결과이다.



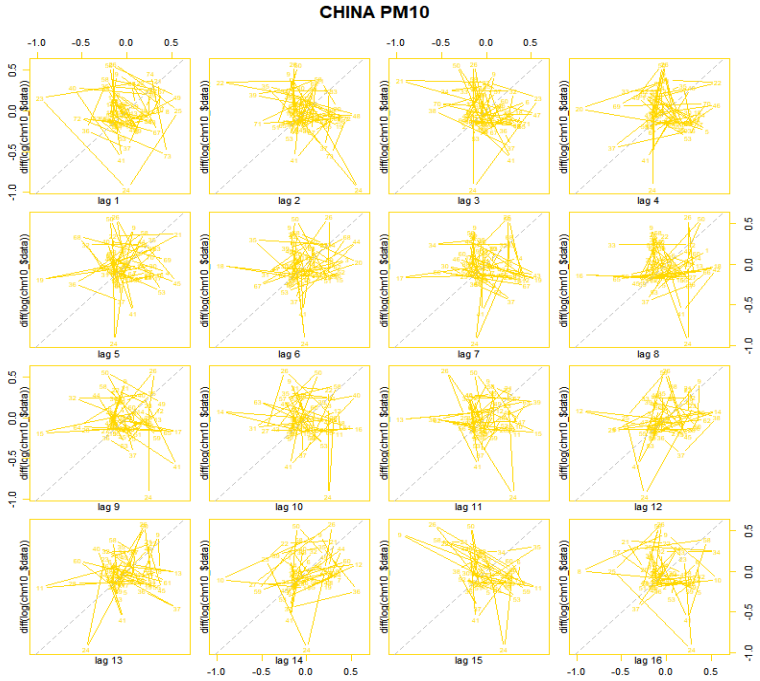
[그림 23] 한국 미세먼지(PM2.5) 데이터의 decomposition plot. 한국 미세먼지(PM2.5) 데이터의 decomposition plot 을 그린 결과이다.

우리의 데이터에서는 중국 미세먼지(PM10), 미세먼지(PM2.5)의 경우 [그림 5], [그림 20], [그림 21]에서 볼 수 있듯이, 평균이 감소하는 trend 를 확인할 수 있었다. 그렇기에 중국 데이터에 한해서 차분을 진행하였다. 차분 후 데이터를 다시 adf.test()를 진행한 결과, 이 역시 정상성을 만족한다고 볼 수 있었다. 한편 한국의 경우, [그림 9], [그림 22], [그림 23]에서 볼 수 있듯이, 계속 감소하는 경향은 보이지 않았기 때문에 따로 차분은 진행하지 않기로 결정했다. 따라서, 중국의 경우 d=1, 한국의 경우 d=0 에 대해 ARIMA 피팅을 진행하였다.

또한, 중국과 한국 미세먼지 모두 계절성을 가지는 것을 앞의 4.1.1.의 그래프에서 볼 수 있었다. 또한, [그림 20], [그림 21]에서 seasonal 파트를 보면 중국 데이터의 seasonal 파트는 여름에 가장 낮고, 겨울에도 일시적 감소가 있으며, 가을에 가장 큰 값을 가지는 변동을 볼 수 있다. [그림 22], [그림 23]에서 seasonal 파트를 보면 한국 데이터는 중국 데이터와 다르게 겨울철 감소하는 경향은 잘 나타나지 않았으며, 여름에 낮고, 겨울에 높은 경향을 보였다. 이를 통해 s=12 임을 유추할 수 있었고, 다른 방법으로도 확인해보고자, periodogram, lag plot, SACF, SPACF 를 그려보았다.

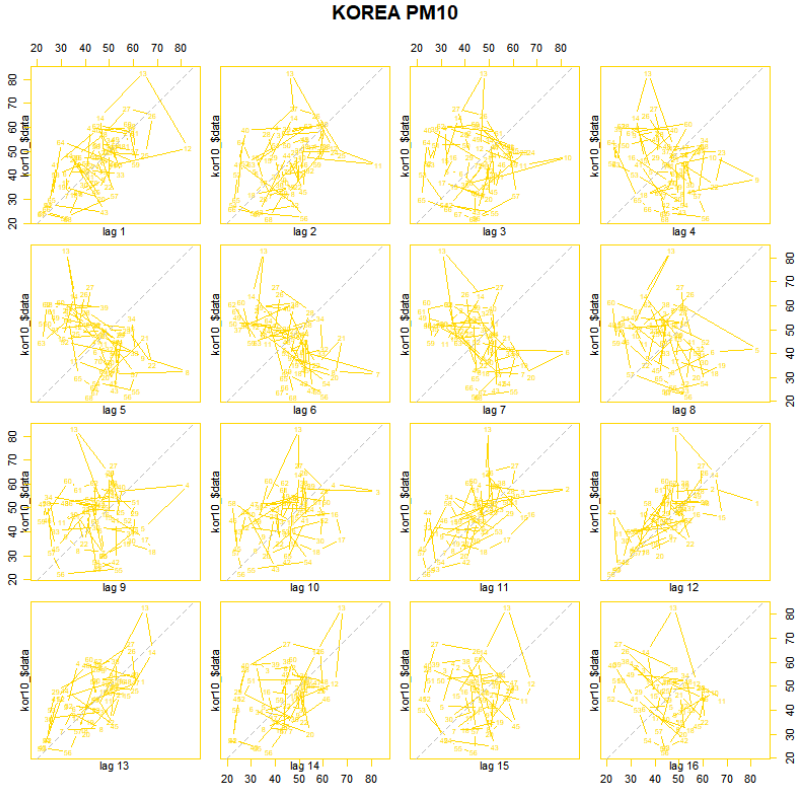


[그림 24] 중국과 한국 미세먼지(PM10), 미세먼지(PM2.5) 데이터의 periodogram. 왼쪽 두 그래프는 위에서부터 각각 중국 log-PM10, PM2.5 periodogram 이며, 오른쪽 두 그래프는 위에서부터 각각 한국 미세먼지(PM10), 미세먼지(PM2.5) periodogram 이다.

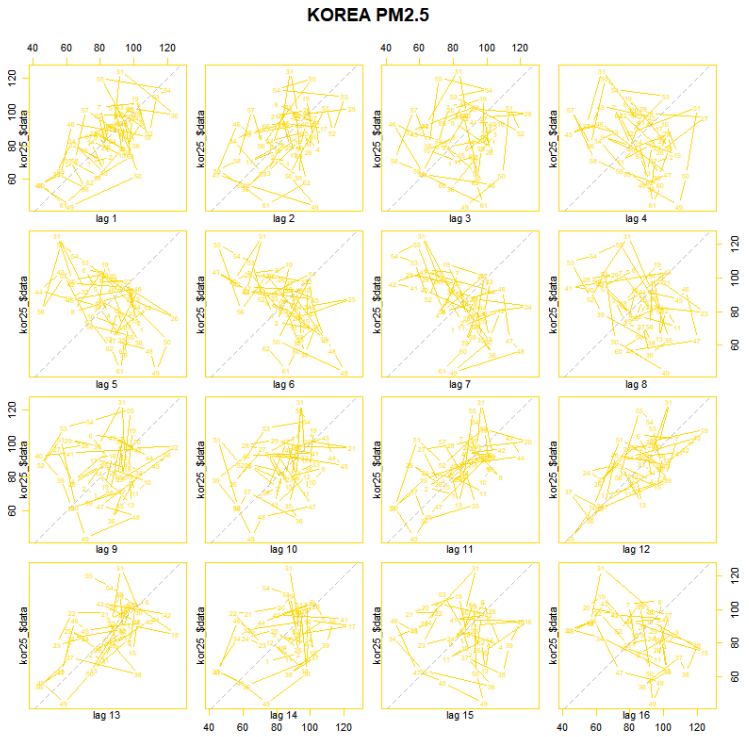


[그림 25] 중국 log-미세먼지(PM10) 데이터의 Lag plot. 중국 log-미세먼지(PM10) 데이터를 차분한 후, Lag plot 을 그린 것이다. 뚜렷한 선형 관계를 보이는 Lag 값을 찾을 수는 없었다.

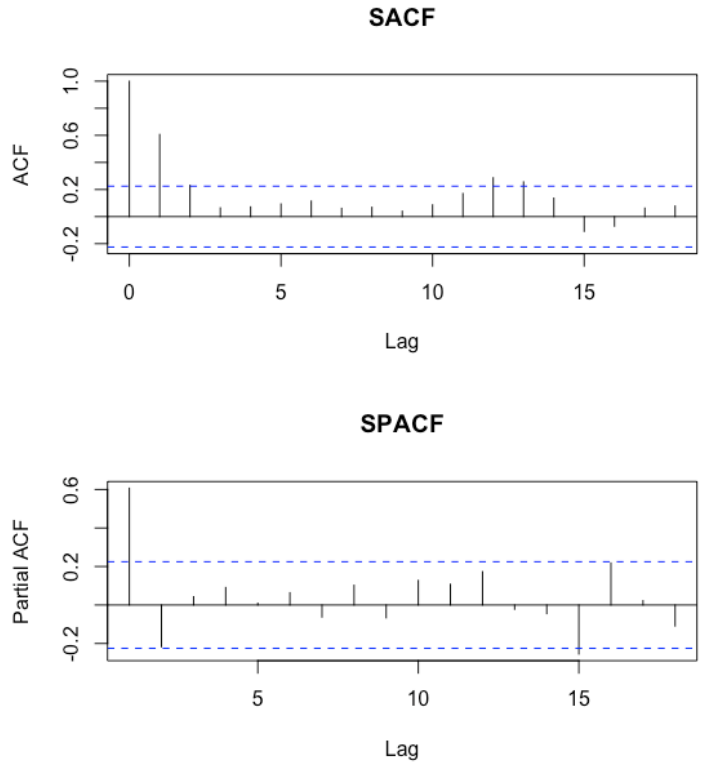
[그림 26] 중국 미세먼지(PM2.5) 데이터의 Lag plot. 중국 미세먼지(PM2.5) 데이터를 차분한 후, Lag plot 을 그린 것이다. 뚜렷한 선형 관계를 보이는 Lag 값을 찾을 수는 없었다.



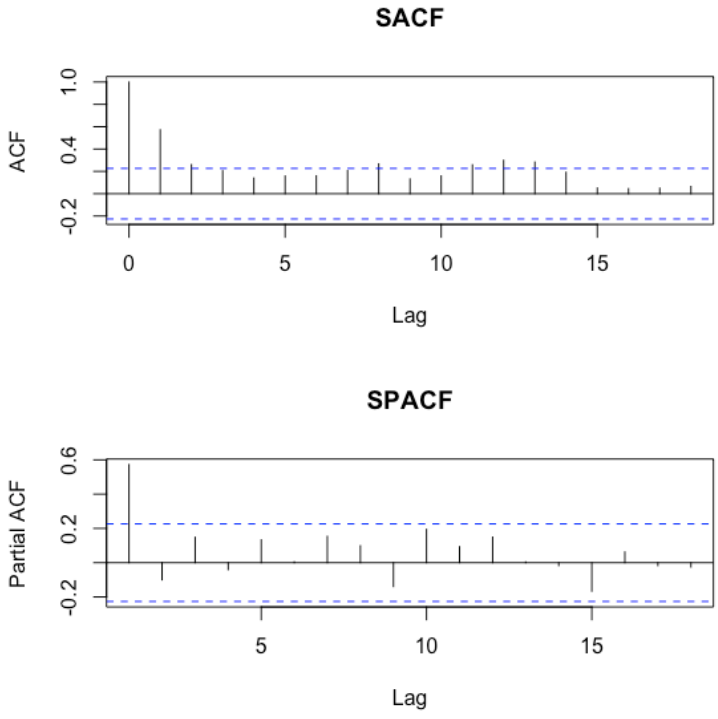
[그림 27] 한국 미세먼지(PM10) 데이터의 Lag plot. 한국 미세먼지(PM10) 데이터의 Lag plot 을 그린 것이다. Lag 12 일 때 뚜렷한 선형 관계를 확인할 수 있었다.



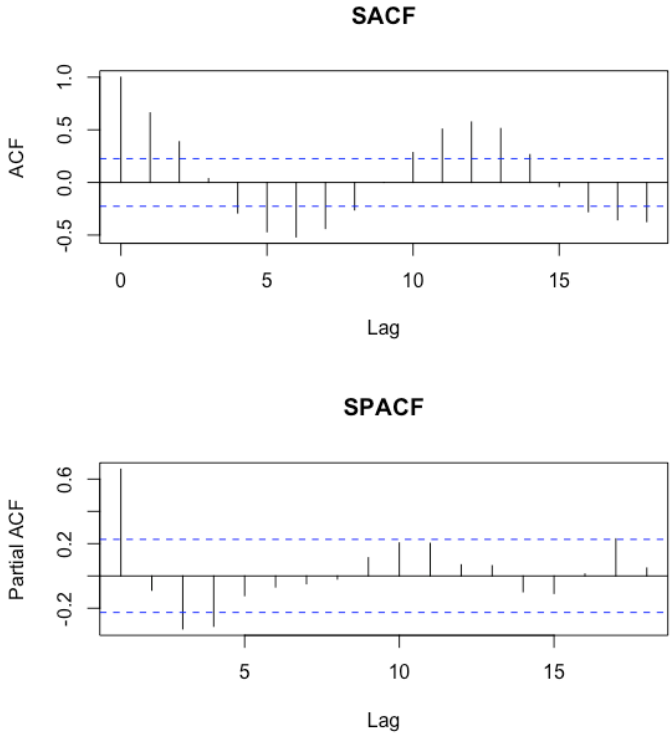
[그림 28] 한국 미세먼지(PM2.5) 데이터의 Lag plot. 한국 미세먼지(PM2.5) 데이터의 Lag plot 을 그린 것이다. Lag 12 일 때 뚜렷한 선형 관계를 확인할 수 있었다.



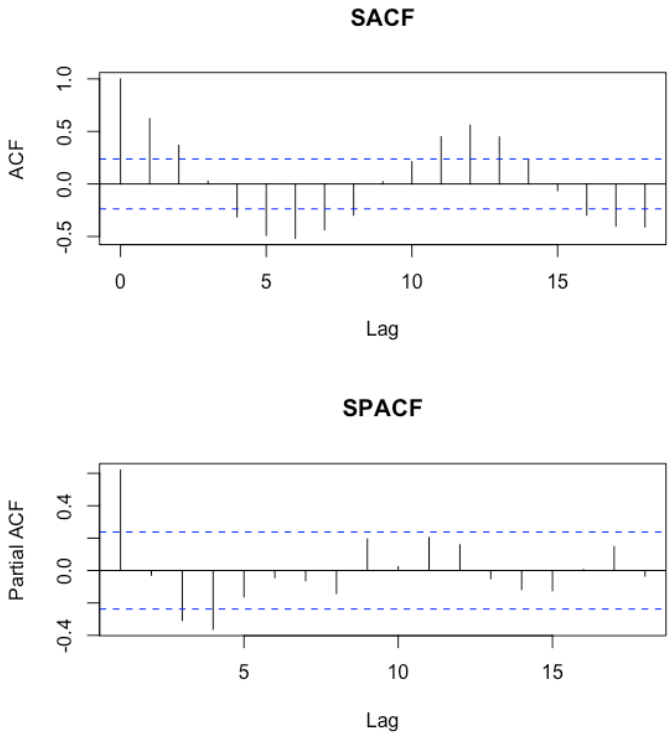
[그림 29] 중국 미세먼지(PM10)의 SACF, SPACF. 중국 미세먼지(PM10) 데이터의 SACF, SPACF 결과이다. h=12 근처에서 커지는 것을 확인할 수 있다.



[그림 30] 중국 미세먼지(PM2.5)의 SACF, SPACF. 중국 미세먼지(PM2.5) 데이터의 SACF, SPACF 결과이다. h=12 근처에서 커지는 것을 확인할 수 있다.



[그림 31] 한국 미세먼지(PM10)의 SACF, SPACF. 한국 미세먼지(PM10) 데이터의 SACF, SPACF 결과이다. h=6,12 근처에서 유의해지고, 파동성을 갖는 것을 확인할 수 있다.



[그림 32]] 한국 미세먼지(PM2.5)의 SACF, SPACF. 한국 미세먼지(PM2.5) 데이터의 SACF, SPACF 결과이다. h=6,12 근처에서 유의해지고, 파동성을 갖는 것을 확인할 수 있다.

먼저, [그림 24]는 periodogram 이다. Spectrum 값이 가장 큰 frequency 의 역수가 period 라고 할 수 있는데, 그래프에서 보면 1/12=0.0833 근처에서 최댓값을 가지는 것을 확인할 수 있었다. 이를 통해 s=12 라고 유추할 수 있었다. 다음은, [그림 25] ~ [그림 28]이다. Lag plot 을 보면, 중국 데이터에서는 아쉽게 뚜렷한 선형관계인 Lag 가 잘 보이진 않았지만, 한국 데이터에서는 Lag 가 12 일 때 뚜렷한 선형 관계를 가짐을 확인할 수 있었다. [그림 29], [그림 30]에서는 중국 데이터의 SACF, SPACF 그래프가 나타나 있는데, h=12 근처에서 커지는 것을 확인하였고, [그림 31], [그림 32]에 있는 한국 데이터의 SACF, SPACF 그래프를 보면 h 는 6,12 근처에서 유의해지고, 파동성을 가짐을 확인할 수 있다. 즉, 계절성을 가짐을 유추할 수 있고, 그 주기가 12 에 가깝다고 생각할 수 있는 충분한 근거가 된다.

이러한 근거들에 의하여 중국은 d=1, D=1, s=12 인 경우, 한국은 d=0, D=1, s=12 로 결정하였고, 이 경우에 대해 ARIMA(p,d,q)(P,D,Q)s 모형에 적합을 하였다.

4.2.2. ARIMA(p,d,q)(P,D,Q)s 모형 적합

	방법	p	d	q	P	D	Q
PM 10	**auto.arima()**	**2**	**1**	**0**	**0**	**1**	**1**
	low AIC	1	1	1	0	1	1
	low BIC	1	1	1	0	1	1
PM 2.5	auto.arima()	0	1	2	2	1	0
	low AIC	**0**	**1**	**2**	**0**	**1**	**1**
	low BIC	**0**	**1**	**2**	**0**	**1**	**1**

[표 2] 중국 Arima(p,d,q)(P,D,Q)s 모형 적합 결과. 중국 미세먼지(PM10, PM2.5) 데이터에 ARIMA(p,d,q)(P,D,Q)s 모형을 적합하였고, auto.arima(), low AIC, low BIC 를 기준으로 한 모형을 뽑았다.

	방법	p	d	q	P	D	Q
PM 10	auto.arima()	0	0	0	2	1	0
	low AIC	**1**	**0**	**1**	**0**	**1**	**1**
	low BIC	**1**	**0**	**1**	**0**	**1**	**1**
PM 2.5	auto.arima()	2	0	0	0	1	1
	low AIC	2	0	0	0	1	1
	low BIC	**1**	**0**	**0**	**0**	**1**	**1**

[표 3] 한국 Arima(p,d,q)(P,D,Q)s 모형 적합 결과. 한국 미세먼지(PM10, PM2.5) 데이터에 ARIMA(p,d,q)(P,D,Q)s 모형을 적합하였고, auto.arima(), low AIC, low BIC 를 기준으로 한 모형을 뽑았다.

이제 4 개의 데이터에 대해, ARIMA(p,d,q)(P,D,Q)s 모형을 적합한 결과는 아래의 [표 2], [표 3]와 같다. 진한 글씨로 표시된 것이 최종적으로 선택된 모형이며, 잔차의 자기상관성 (Q-통계량), 잔차의 정규성, 모수의 유의성을 고려하여 선택하였다. 이에 대한 설명을 다음 절에서 이어서 진행한다.

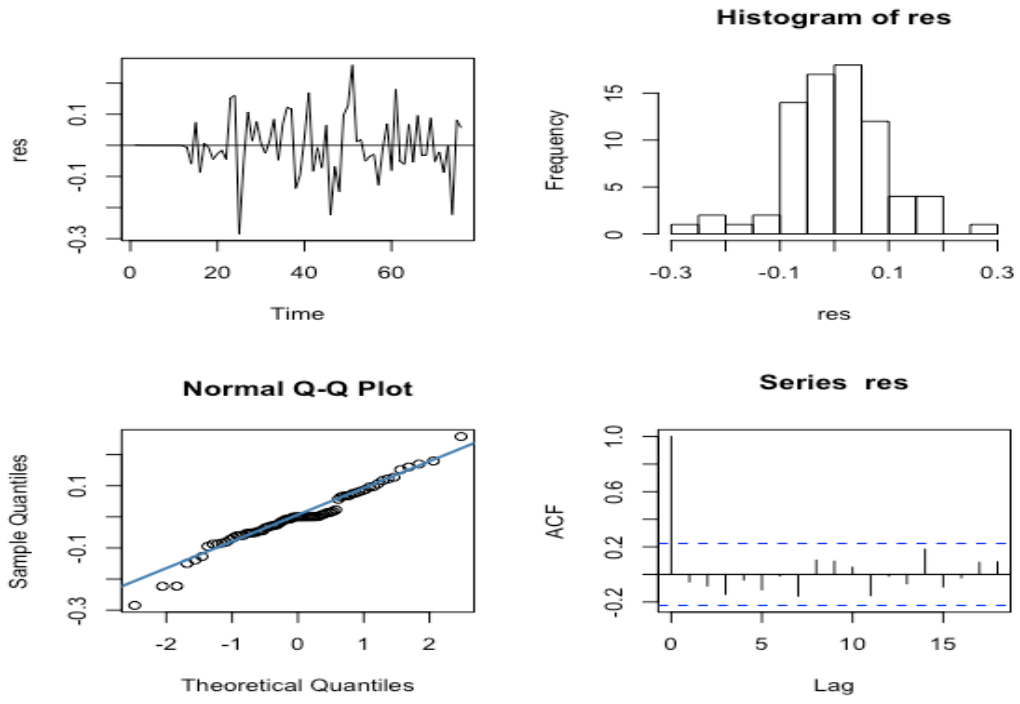
4.2.3. 모형 진단을 통한 최종 모형 결정

먼저, 중국 미세먼지(PM10) 데이터의 최종 모형은 ARIMA(2,1,0)(0,1,1)[12]로 결정되었으며, 잔차들의 plot, 히스토그램, qqplot, SACF 그래프는 아래와 같다.

	p-value	conclusion
D-F(d=1)	0.01	stationary
Q	0.6286	independent

coef.	estimates	p-value
ϕ_1	-0.2365	0.0566
ϕ_2	-0.3928	0.0018
Θ_1	-0.7770	0.0021

[표 4] 중국 미세먼지(PM10) 데이터의 최종 모형의 ADF test 결과, Q 통계량 및 추정된 모수의 값과 p-value. 정상성을 만족하고, 잔차들은 자기상관성을 가지지 않았으며, 모든 모수가 유의하였다.



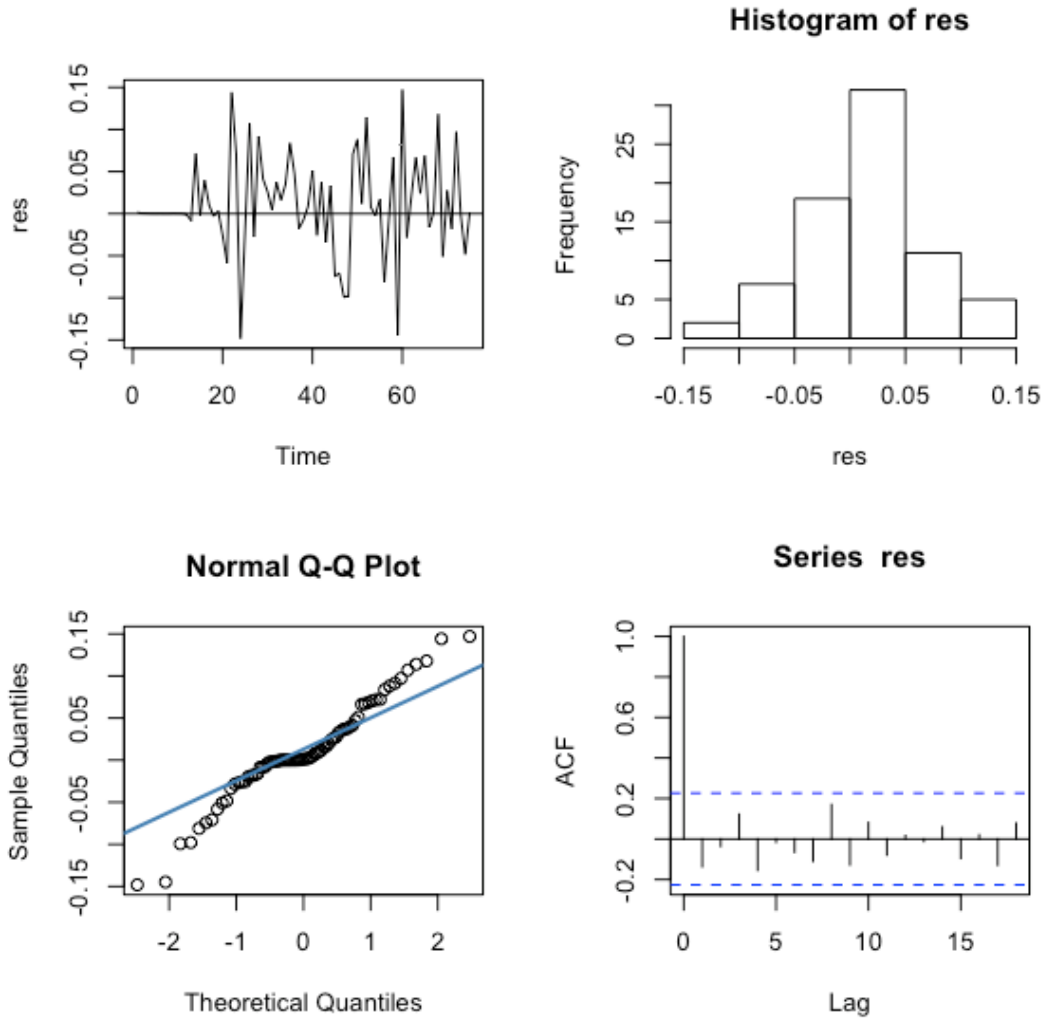
[그림 33] 중국 미세먼지(PM10) 데이터의 최종 모형의 잔차들의 plot, 히스토그램, SACF, qqplot. 왼쪽 위에서부터 시계 방향으로 중국 미세먼지(PM10) 데이터의 최종 모형인 ARIMA(2,1,0)(0,1,1)[12]의 잔차 plot, 히스토그램, SACF 그래프, qqplot 이다. 잔차들이 정규성을 만족하며, 자기상관성을 가지지 않았다.

다음은, 중국 미세먼지(PM2.5) 데이터의 최종 모형이다. 최종 모형은 ARIMA(0,1,2)(0,1,1)[12]로 결정되었으며, 잔차들의 plot, 히스토그램, qqplot, SACF 그래프는 아래와 같다.

	p-value	conclusion
D-F(d=1)	0.01	stationary
Q	0.2259	independent

coef.	estimates	p-value
θ_1	-0.2661	0.0778
θ_2	-0.7338	3.314e-07
Θ_1	-0.9964	0.3229

[표 5] 중국 미세먼지(PM2.5) 데이터의 최종 모형의 ADF test 결과, Q 통계량 및 추정된 모수의 값과 p-value. 정상성을 만족하고, 잔차들은 자기상관성을 가지지 않았으며, θ_1을 제외한 모수가 유의하였다.



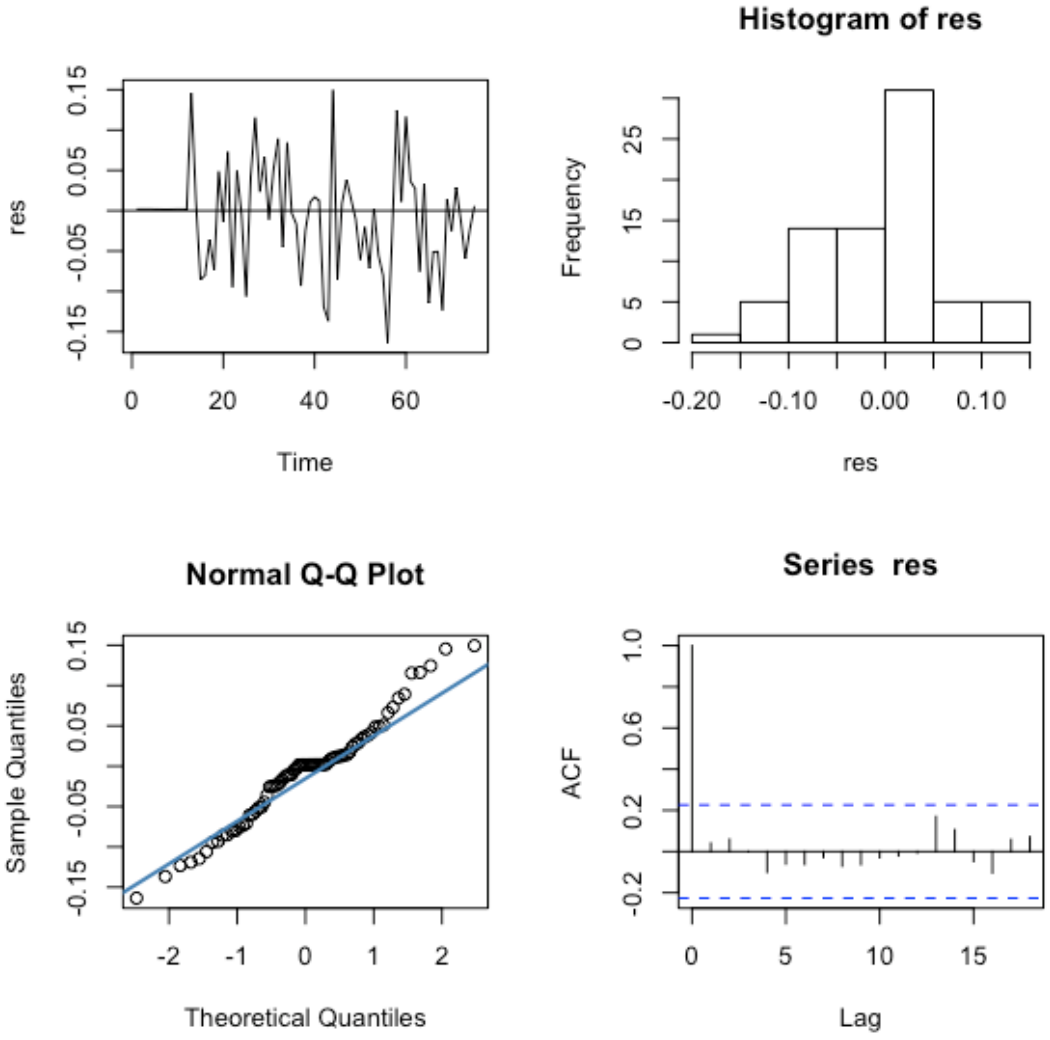
[그림 34] 중국 미세먼지(PM2.5) 데이터의 최종 모형의 잔차들의 plot, 히스토그램, SACF, qqplot. 왼쪽 위에서부터 시계 방향으로 중국 미세먼지(PM2.5) 데이터의 최종 모형인 ARIMA(0,1,2)(0,1,1)[12]의 잔차 plot, 히스토그램, SACF 그래프, qqplot 이다. 잔차들이 정규성은 잘 만족되지 않았으며, 자기상관성을 가지지 않았다.

한국 미세먼지(PM10) 데이터의 최종 모형은 ARIMA(1,0,1)(0,1,1)[12]로 결정되었으며, 잔차들의 plot, 히스토그램, qqplot, SACF 그래프는 아래와 같다.

	p-value	conclusion
D-F(d=1)	0.01	stationary
Q	0.7185	independent

coef.	estimates	p-value
ϕ_1	0.9966	0
θ_1	-0.8416	3.2817e-10
Θ_1	-0.9629	0.0134

[표 6] 한국 미세먼지(PM10) 데이터의 최종 모형의 ADF test 결과, Q 통계량 및 추정된 모수의 값과 p-value. 정상성을 만족하고, 잔차들은 자기상관성을 가지지 않았으며, 모든 모수가 유의하였다.

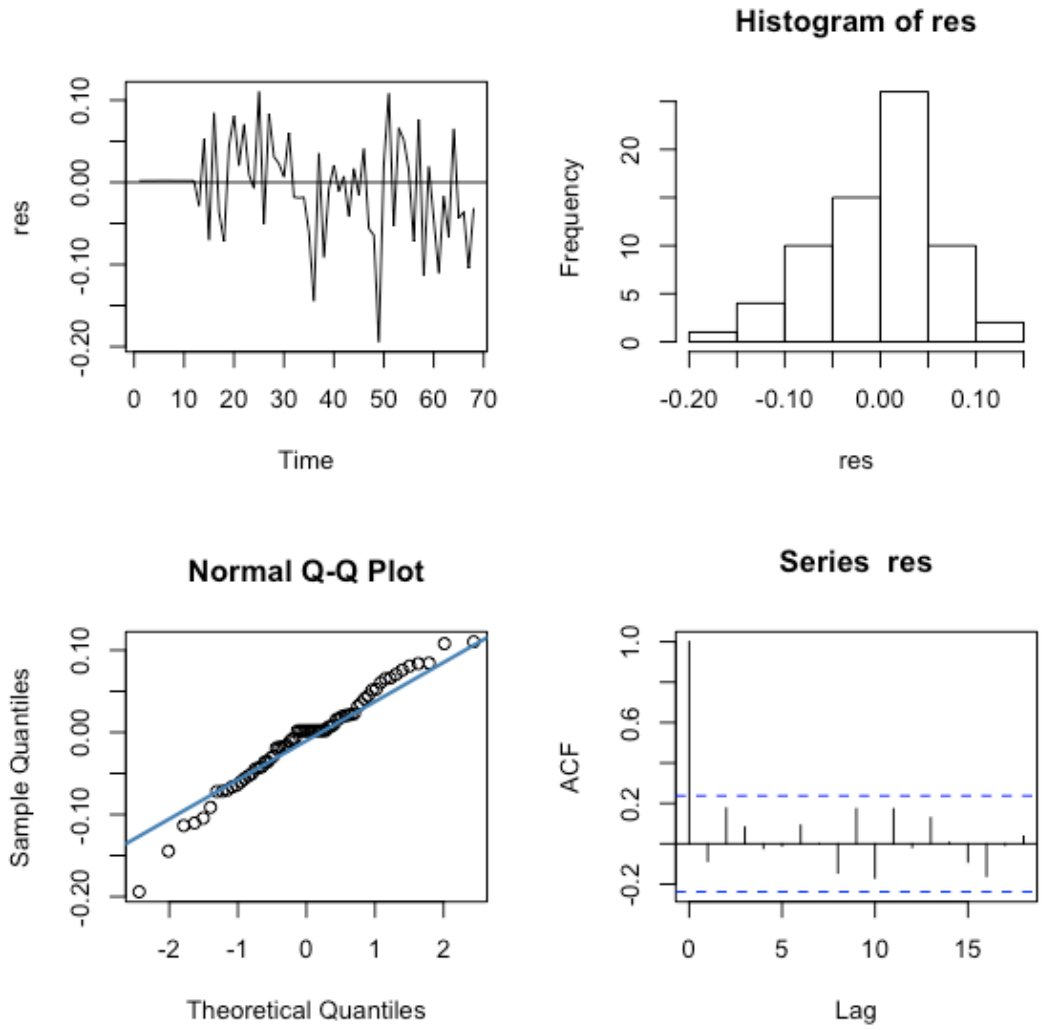


[그림 35] 한국 미세먼지(PM10) 데이터의 최종 모형의 잔차들의 plot, 히스토그램, SACF, qqplot. 왼쪽 위에서부터 시계 방향으로 한국 미세먼지(PM10) 데이터의 최종 모형인 ARIMA(1,0,1)(0,1,1)[12]의 잔차 plot, 히스토그램, SACF 그래프, qqplot 이다. 잔차들이 정규성을 어느 정도 만족했으며, 자기상관성을 가지지 않았다.

마지막으로 한국 미세먼지(PM2.5) 데이터의 최종 모형은 ARIMA(1,0,0)(0,1,1)[12]로 결정되었으며, 잔차들의 plot, 히스토그램, qqplot, SACF 그래프는 아래와 같다.

	p-value	conclusion	coef.	estimates	p-value
D-F(d=1)	0.01	stationary	ϕ_1	0.3397	0.0175
Q	0.4707	independent	Θ_1	-0.6243	0.0059

[표 7] 한국 미세먼지(PM2.5) 데이터의 최종 모형의 ADF test 결과, Q 통계량 및 추정된 모수의 값과 p-value. 정상성을 만족하고, 잔차들은 자기상관성을 가지지 않았으며, 모든 모수가 유의하였다.

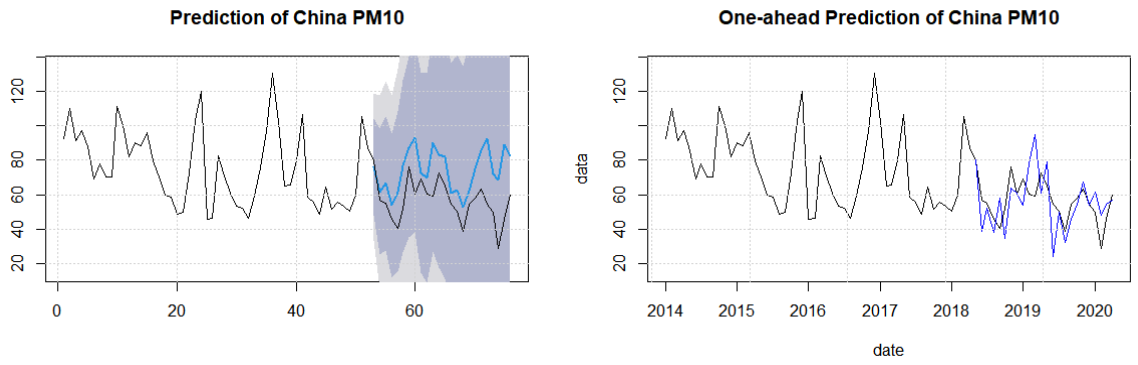


[그림 36] 한국 미세먼지(PM2.5) 데이터의 최종 모형의 잔차들의 plot, 히스토그램, SACF, qqplot. 왼쪽 위에서부터 시계 방향으로 한국 미세먼지(PM2.5) 데이터의 최종 모형인 ARIMA(1,0,0)(0,1,1)[12]의 잔차 plot, 히스토그램, SACF 그래프, qqplot 이다. 잔차들이 정규성을 어느 정도 만족했으며, 자기상관성을 가지지 않았다.

4.3. 예측

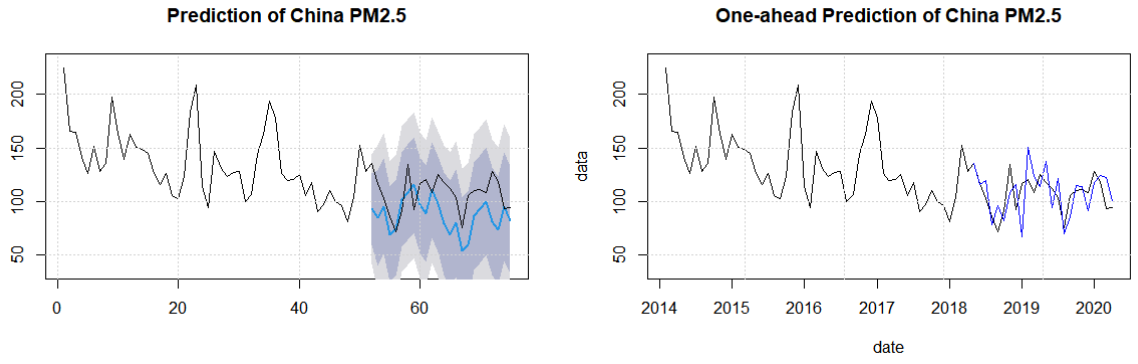
4.3.1. Forecast 함수를 이용한 예측

앞서 4.2.3 에서 구한 최종 모형을 이용하여 최근 2 년 데이터를 예측하는 prediction 과 one-ahead prediction 를 한 결과는 다음과 같다.



[그림 37] 중국 미세먼지(PM10) 데이터의 prediction 및 one-ahead prediction 그래프. 검은색 값은 실제 값이며, 파란색 값은 prediction 된 값을 나타낸다.

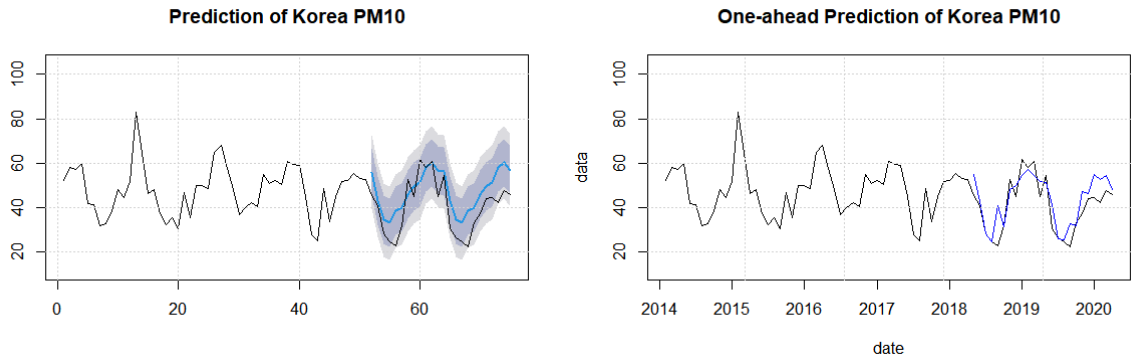
[그림 37]은 중국 미세먼지(PM10) 데이터의 최종 모형인 ARIMA(2,1,0)(0,1,1)[12]을 이용하여, prediction 을 진행한 결과이다. 2 년 이전의 데이터로 예측 시 어느 정도 경향성을 잘 따라갔으나, 예측값이 실제값보다 더 높게 나타났다. One-ahead prediction 결과는 어느 정도 실제 값을 잘 예측했다고 볼 수 있다.



[그림 38] 중국 미세먼지(PM2.5) 데이터의 prediction 및 one-ahead prediction 그래프. 검은색 값은 실제 값이며, 파란색 값은 prediction 된 값을 나타낸다.

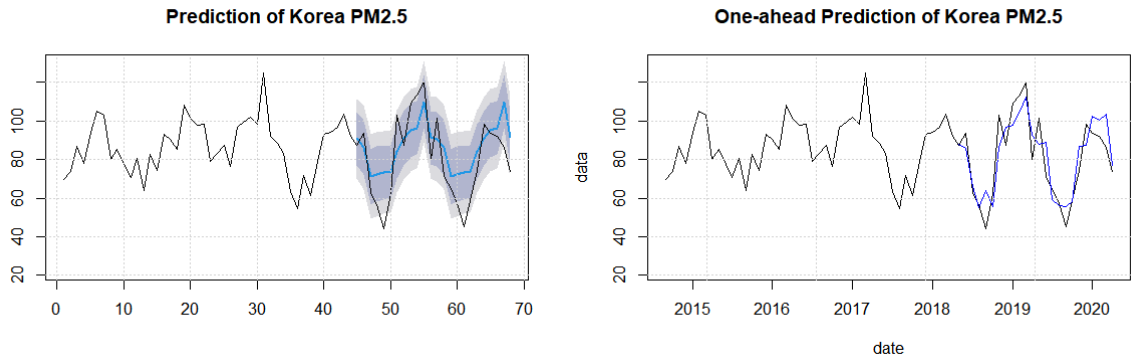
[그림 38]은 중국 미세먼지(PM 2.5) 데이터의 최종 모형인 ARIMA(0,1,2)(0,1,1)[12]을 이용하여, prediction 을 진행한 결과이다. 중국 미세먼지(PM10)와 유사하게, 2 년 이전의 데이터로 예측 시 어느 정도 경향성을 잘 따라갔으며, 예측값이 실제값보다 더 낮게 나타났다. 또한, One-ahead prediction 예측 결과도 어느 정도 실제 값을 잘 예측한다고 볼 수 있다.

다음은 한국 미세먼지(PM10), 미세먼지(PM2.5) 데이터의 최종 모형을 이용하여 각각 예측을 한 결과이다.



[그림 39] 한국 미세먼지(PM10) 데이터의 prediction 및 one-ahead prediction 그래프. 검은색 값은 실제 값이며, 파란색 값은 prediction 된 값을 나타낸다.

[그림 39]는 한국 미세먼지(PM10) 데이터의 최종 모형인 ARIMA(1,0,1)(0,1,1)[12]을 이용하여 예측한 결과이다. 중국 미세먼지(PM10) prediction 과 유사하게, 2 년 이전의 데이터로 예측 시 경향성은 잘 따라갔지만, 예측값이 실제값보다 더 높게 나타났다. One-ahead prediction 결과는 어느 정도 실제 값을 잘 예측했다고 볼 수 있다.



[그림 40] 한국 미세먼지(PM2.5) 데이터의 prediction 및 one-ahead prediction 그래프. 검은색 값은 실제 값이며, 파란색 값은 prediction 된 값을 나타낸다.

[그림 40]은 한국 미세먼지(PM2.5) 데이터의 최종 모형인 ARIMA(1,0,0)(0,1,1)[12]를 이용하여 예측한 결과이다. 2 년 이전의 데이터로 예측 시 어느정도 경향성을 잘 따라갔으며, 8-9 월 부근에서 실제값이 예측값의 95% 신뢰구간 이하로 미세먼지 관측치가 크게 낮아지는 특징이 보였다. 2018 년 4 월을 기준, 연 평균 미세먼지 농도를 비교해본 결과 87.03 에서 80.76 로 평균 미세먼지 농도가 감소하는 현상을 확인할 수 있었다. 또한, One-ahead prediction prediction 그래프를 보면 어느 정도 실제 값을 잘 예측한다고 볼 수 있다.

5. Conclusion

본 연구에서 AIC 와 BIC 를 최소화하면서 동시에 잔차들이 조건 및 모수들의 유의성까지 고려하여 중국의 미세먼지(PM10) 데이터는 ARIMA(2,1,0)(0,1,1)[12] 모형, 중국의 미세먼지(PM2.5) 데이터는 ARIMA(0,1,2)(0,1,1)[12] 모형, 한국의 미세먼지(PM10) 데이터는 ARIMA(1,0,1)(0,1,1)[12] 모형, 한국의 미세먼지(PM2.5) 데이터는 ARIMA(1,0,0)(0,1,1)[12] 모형을 적합할 수 있었다. 각각의 모형들로 one-ahead prediction 을 진행하였을 때 예측값이 실제값을 잘 예측하여 적합한 모형이라고 생각할 수 있다.

본 연구에서 확인할 수 있던 몇 가지 사실들이 있다. 미세먼지(PM10)과 미세먼지(PM2.5)의 농도는 유사한 경향성을 갖는다는 전문가들의 주장이 있었다. 각 데이터들을 plotting 했을 때 두 미세먼지의 농도는 정말 유사한 경향성을 가지는 것을 확인할 수 있었다. 미세먼지(PM10)과 미세먼지(PM2.5)는 주로 같이 발생하여 확산이 되기 때문에 농도의 경향성이 유사하다고 생각해볼 수 있었다. 한국과 중국의 미세먼지 농도의 경향성도 유사한 것을 확인할 수 있었다. 데이터의 plot 뿐만 아니라 다양한 근거를 바탕으로 미세먼지의 주기는 12 개월로 볼 수 있었다.

적합한 모형을 바탕으로 데이터들을 예측하였을 때 대부분 잘 예측하였지만 한국의 미세먼지(PM2.5)의 경우 잘 맞지 않는 부분이 있었다. 실제값이 예측값보다 확연하게 낮게 나타나는 지점이 존재했다. 주로 여름철의 실제값이 예측값보다 낮은 값을 갖고 있었다. 해당 결과의 원인을 다양한 측면에서 분석해보았다. 첫째, 한국에서는 미세먼지 저감을 위해 다양한 노력을 하고 있다. 2017 년 9 월 한국에서 미세먼지 관리 종합대책이 수립되었으며 2018 년 1 월에는 비상, 상시 미세먼지 관리 강화대책이 수립되었고 2019 년 2 월에는 미세먼지 특별법이 시행되기도 하였다. 둘째, 미세먼지 농도는 국내, 외의 영향과 바람이나 대기 안정도, 습도 등의 복합적인 영향으로 결정된다. 두 가지 원인들을 고려하였을 때 한국의 미세먼지 저감 조치는 한국의 미세먼지 농도를 일정 부분은 감소시키는 데 영향을 주었지만 편서풍이 강해져서 중국에서 넘어온 황사가 자주 발생하는 시기인 봄, 겨울철 미세먼지에는 다양한 저감 노력들이 영향을 없었기 때문에 예측값과 실제값이 큰 차이가 없었다. 하지만 편서풍의 영향이 적은 8 월과 9 월에는 미세먼지 저감 조치의 효과로 미세먼지 농도가 줄어든 결과를 갖게 된 것으로 생각할 수 있다.

본 연구에는 몇 가지 한계점이 존재한다. 첫째, 결측치를 선형적으로 처리했다는 점이다. 시계열 자료는 이전의 데이터에 영향을 받을 수 있기 때문에 해당 방법을 사용했지만 선형적으로만 채웠기 때문에 한계가 있을 수 있다. 하지만 결측치를 채운 부분이 전체 데이터 개수에 비해 적었고 분석은 월별 데이터의 평균을 이용했기 때문에 결과에 큰 영향을 주지는 않았을 것으로 판단한다. 둘째, 분석에 있어 월별 평균 값을 이용하였다는 점이다. 일별 자료에서는 시계열 모형 적합이 적절하지 않다는 분석 결과가 있어 부득이 하게 월별 평균 값을 이용하여 시계열 모형 적합을 실시하였다. 일별 자료에 비해 데이터 개수가 적어 적합 결과가 정확하지 않을 수 있다. 하지만 해석에 있어서 일별 자료보다는 월별 자료가 수월하며 주기가 12 개월인 미세먼지의 특성상 월별 자료를 바탕으로 분석을 하더라도 큰 문제는 없을 것이라고 생각한다. 마지막으로 한국과 중국 미세먼지 농도 경향성의 유사성은 확인할 수

있었지만 많은 사람들이 생각하는 것처럼 중국이 한국 미세먼지 농도의 원인이지는 파악할 수 없었다. 본 연구에서는 한국과 중국의 미세먼지 데이터를 적합하기 위한 다양한 근거를 바탕으로 시계열 모형을 찾고 적절한 모형을 바탕으로 예측하여 모형의 적합성을 확인하는 것에 초점이 맞추어져 있다. 따라서 중국과 한국의 미세먼지의 상관관계는 분석하지 않았으며 직접적인 원인이 되는 것을 확인하기 위한 분석을 따로 진행하지 않았다. 해당 결과를 확인하기 위해서는 회귀분석과 같은 다른 방법을 이용하여 분석을 진행해야 한다고 생각한다.

6. Reference

[1] 이상열(2012),『시계열분석 이론 및 SAS실습』, 자유아카데미