한국농림기상학회지, 제22권 제3호(2020) (pISSN 1229-5671, eISSN 2288-1859)
Korean Journal of Agricultural and Forest Meteorology, Vol. 22, No. 3, (2020), pp. 171~175
DOI: 10.5532/KJAFM.2020.22.3.171

DATA

시군구별 대기오염 및 기상 데이터

윤성도[1] · 김승규[2*]
[1]미시시피주립대학교 농업경제학과, [2]경북대학교 농업경제학과
(2020년 8월 10일 접수; 2020년 8월 26일 수정; 2020년 9월 1일 수락)

Air Pollution and Weather Data by Si-Gun-Gu in South Korea

Seong Do Yun[1], Seung Gyu Kim[2*]
[1]*Department of Agricultural Economics, Mississippi State University*
[2]*Department of Agricultural Economics, Kyungpook National University*
(Received August 10, 2020; Revised August 26, 2020; Accepted September 1, 2020)

ABSTRACT

Studies in socioeconomic impacts of air pollution are inevitable to merge data of the air pollutant density, weather, and socioeconomic variables. Due to their spatiotemporal disparities in units, to combine these data are time and effort consuming generically. The data described in this article aims to provide the major variables of air pollution and weather at the Si-Gun-Gu level to meet the data needs from social science. The latest (August 2020) data distributed are the balanced panel of 250 Si-Gun-Gu in South Korea for 2001-2018. The weather variables in this data are directly applicable to other social science topics, which are not limited to air pollution research.

Key words: **Air pollution, Weather, Si-Gun-Gu, Korea**

I. 배경 및 요약

대기오염의 사회경제적 영향은 세계적으로 활발히 진행되고 있는 학제간 연구(multidisciplinary studies) 주제의 한 분야이다. 2012년을 전후하여 높아진 미세먼지에 대한 관심을 반영하여, 우리나라에서도 대기오염에 관한 다양한 연구들이 진행되었다. 학제간 연구에서 흔히 직면하게 되는 어려움 중의 하나는 자연과학 또는 공학에서 사용하는 데이터와 인문사회과학에서 사용하는 데이터의 분석단위가 다르다는 점을 들 수 있다. 예를 들어, 대기오염 물질의 경우 각 관측소에서 시간대별로 측정한 자료가 존재하지만, 주요 인구 및 사회경제 관련 변수들은 시군구나 시도와 같은 행정구역 단위로 연도별 집계가 대부분을 이룬다.

시간 및 공간 단위가 불일치하는 자료의 사용은 연구 신뢰성에 영향을 미칠 수 있다. Yun and Gramig (2019)에서 지적한 바와 같이 데이터를 더 넓은 면적 단위로 평균하게 되면 집계편향(aggregation bias)이 발생하고, 이 집계편향이 통계 및 수리적 분석 결과에 영향을 미친다. 우리나라 대기오염의 사회경제적 영향에 관한 연구에서 널리 쓰이는 자료는 16개 광역시도 및 세종시를 기준으로 한 PM10 및 PM2.5 연간 평균 농도 자료이다. 이들 자료는 국가통계포털에서 제공되는 사회경제적 변인들과 함께 손쉽게 사용할 수 있기 때문이다. 또한, 사회과학자들이 시간 단위로 집계되는 관측소 수준의 데이터 존재를 인지한다 하더라도

* Corresponding Author : Seung Gyu Kim
(sgkimwin@knu.ac.kr)

이를 행정구역 단위로 변환하는 방법에 대한 훈련 부족으로 사용이 쉽지 않다. 그러나 광역시도 수준의 관측치로 분석할 경우 통계적으로 유의미한 결과를 도출하기 어렵다. 따라서 광역시도 수준보다는 다양한 사회경제적 변수가 가용한 시군구 수준에서 대기오염 및 기상변수 데이터를 구축하여 분석하는 것이 집계편향을 완화하고 보다 유의성 높은 추정치를 얻는데 도움이 될 것이다.

본고에서 기술하는 데이터는 대기오염의 사회경제적 영향을 시군구 단위로 분석할 경우 필요한 주요 대기오염 물질 및 기상 자료를 담고 있다. 2017년 기준 250개 시군구를 기준으로 2001년부터 2018년까지의 연간 대기오염 물질 농도 및 기상 변수들을 아래 세 가지 연구 수요의 충족을 위해 가공하였다. 첫째, 사회경제적 분석에서 주로 사용하는 대기오염 및 기상 변수를 제공한다. 둘째, 사회경제적 변수와 대기오염 및 기상 변수가 시군구와 연도를 기준으로 병합할 수 있는 데이터를 생성한다. 마지막으로, 비교적 간단한 모형을 적용하되 자연과학 및 공학에서도 널리 인정되는 방법론을 사용하여 데이터를 구축한다.

본고의 데이터는 대기오염 인지에 따른 경제적 행위의 변화를 연구한 Yun and Kim(2020b)에서 적용된 바가 있다. Yun and Kim(2020b)은 한국노동연구원의 노동패널과 본 데이터를 병합하여 사용하였는데, 이는 본 데이터를 다양한 사회경제적 분석에 사용할 수 있음을 보여주는 좋은 예이다. 또한 본 데이터에서 제공하는 기상 변수는 Hsiang(2016)에서 개념화한 기후계량경제모형(climate econometrics)의 다양한 모형에서 공통적으로 사용하는 변수들을 모두 포함하였다. 이는 Schlenker and Roberts(2009)의 crop yield response function이나 Deschênes and Greenstone(2007)의 panel approach를 활용한 기상 및 기후변화의 사회 경제적 분석 모형을 우리나라에 적용시 필요한 대부분의 변수를 본 데이터가 제공하고 있음을 의미한다.

II. 방 법

우리나라 행정구역은 정치·경제적 목적에 의해 지속적으로 통합 또는 분리되어 공간적 단위가 변경되어 왔다. 사회과학에서 널리 사용되는 패널자료 형태로 데이터를 구축하기 위해 지오서비스(Geoservice, 2020)에서 제공하는 2017년 3월 대한민국 최신 행정구역 shapefile을 기준으로 삼았다. 이를 통해, 총 250개의 시군구 centroid 좌표를 추출하여 해당지역의 대기오염 및 기상 변수들을 추정하였다. Fig. 1은 250개 시군구 행정구역과 대기오염 물질과 기상 데이터의 원자료가 수집된 측정소의 위치를 보여준다.

대기오염 물질의 원자료는 한 시간 단위로 Fig. 1 좌측에 표시된 측정소들에서 수집된 에어코리아(AirKorea, 2020)의 최종확정 측정자료이다. 이들 원자료를 시간대 별로 정규크리깅(Ordinary Kriging)을 통해 250개 시군구의 centroid 좌표에서의 추정값을 구하였다. 울릉군, 제주시 및 서귀포시의 경우 정규크리깅의 가정(Beguería *et al*., 2016)을 만족하지 않으므로 해당지역의 측정소 원자료를 그대로 사용하였다. 정규크리깅의 추정에는 최우추정법(maximum likelihood estimation)(Cressie, 2015)을 적용하였다. 원자료에서 사용된 대기오염 물질 농도 변수들은 국제적으로 통용되는 대기

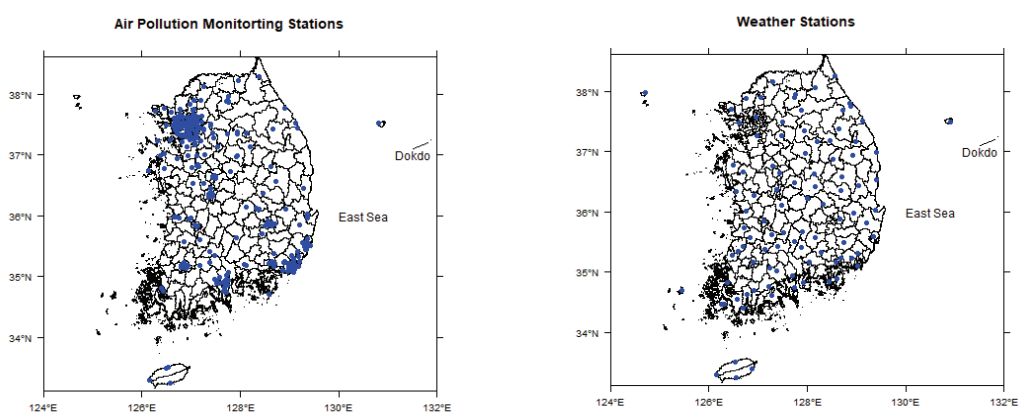

Fig. 1. Weather Stations and Air Pollution Monitoring Stations.

오염 농도 측정값인 PM10, CO, NO_2, SO_2, O_3이다. 이들 데이터는 다음 장에서 설명된 변수의 정의에 따라 시간별 추정치를 연간 단위로 계산하여 다양한 변수의 정의에 따라 변환하였다. PM2.5의 경우는 원자료에서 2015년부터 제공하고 있기 때문에, 현재는 포함 되어 있지 않으나 추후 추가할 계획이다.

기상 데이터의 원자료는 일별 단위로 Fig. 1 우측에 표시된 측정소들에서 수집된 기상청 기상자료개방포털(KMA, 2020)의 기상관측 자료이다. 기상관측소는 250개 시군구 전역에 걸쳐 고르게 분포하고 있기 때문에 centroid 좌표에서 Euclidian 거리로 가장 가까운 관측소의 값을 사용하였다. 기상 데이터의 원자료 중 사회과학에서 가장 널리 사용되는 기온, 강수량, 풍속 및 풍향, 습도, 일조시간 변수를 대기오염 물질과 마찬가지로 다음 장에서 설명된 변수의 정의에 따라 일간 추정치를 연간 단위로 계산하였다. 대기오염 자료와 마찬가지 이유로 울릉군의 경우 해당 지역의 측정소 자료를 그대로 적용하였다.

III. 자료 및 사용방법

본고의 데이터는 저자가 운영하는 Github 사이트(https://github.com/ysd2004/AirKorea, Yun and Kim, 2020a)에서 접근가능하며 가용한 변수에 대한 상세설명

Table 1. Variable Description

Variable	Description (KOR)	Description (ENG)
SIG_CD	시군구코드	Si-Gun-Gu Code
SIG_ENG	영문 시군구명	Si-Gun-Gu Name
year	연도	Year
COavg	연평균 일산화탄소(CO) 농도(ppm)	Yearly CO average (ppm)
NO_2avg	연평균 이산화질소(NO_2) 농도(ppm)	Yearly NO_2 average (ppm)
O_3avg	연평균 오존 (O_3) 농도(ppm)	Yearly O_3 average (ppm)
PM10avg	연평균 PM10 농도(㎍/m^3)	Yearly PM10 average (㎍/m^3)
SO_2avg	연평균 아황산가스(SO_2) 농도(ppm)	Yearly SO_2 average (ppm)
PM10freq	한국기준 PM10>80(나쁨, 매우나쁨)인 날수	Day counts of PM10>80 (unhealthy and very unhealthy in the Korean Standard)
PM10freqUS	미국기준 PM10>154(민감군 나쁨, 나쁨, 매우나쁨)인 날수	Day counts of PM10>154 (S-groups unhealthy, unhealthy and very unhealthy in the US Standard)
PM10xtrm	한국기준 PM10>150(매우나쁨)인 날수	Day counts of PM10>150 (very unhealthy in Korean the Standard)
PM10xtrmUS	미국기준 PM10>354(매우나쁨)인 날수	Day counts of PM10>354 (very unhealthy in the US Standard)
PM10hrs	한국기준 PM10>80(매우 나쁨) 노출시간	Length of hours with PM10>80 (very unhealthy in the Korean Standard)
avghrs	한국기준 PM10>80(매우 나쁨) 일간 평균 노출시간	Daily average exposure hours with PM10> 80(very unhealthy in the Korean Standard)
tavg	연평균기온(°C)	Yearly average temperature (°C)
tmin	연평균 최저기온(°C)	Yearly average minimum temperature(°C)
tmax	연평균 최고기온(°C)	Yearly average maximum temperature (°C)
ppt	연강수량(mm)	Yearly total precipitation (mm)
wmax	연평균 최대풍속(m/s)	Yearly average of the maximim wind speeds (m/s)
wmaxd	연평균 최대풍속 방향(°)	Yearly average of the maximim wind speed direction (°)
wavg	연평균 풍속(m/s)	Yearly average of the mean wind speed (m/s)
huminavg	연평균 상대습도(%)	Yearly average of the relative humidity (%)
sunsum	연간 일조시간(hrs)	Yearly total shunshine hours (hrs)

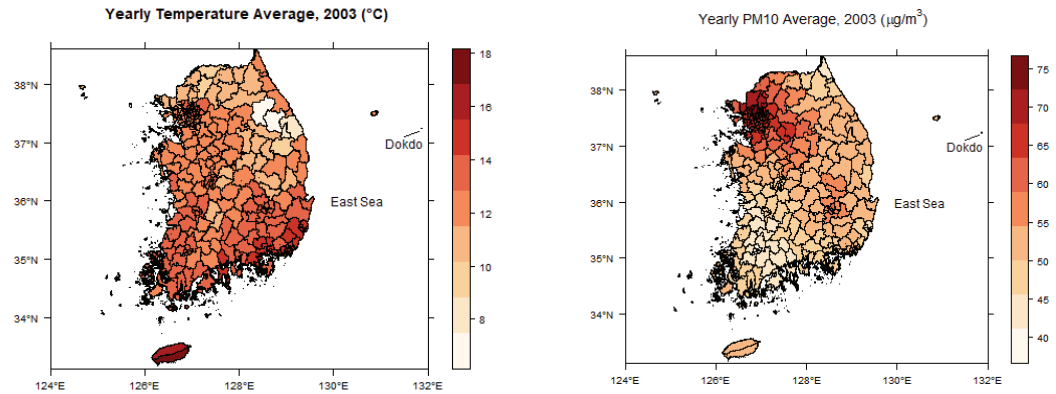

Fig. 2. Example Data Plots: Yearly Temperature and PM10 Averages, 2003.

을 참조할 수 있다. 2020년 8월 기준, 본 데이터의 버전은 1.0.0이며, 패널형태의 자료를 CSV 형식으로 다운로드 가능하다. 본 데이터에 포함된 변수는 Table 1과 같다. Table 1의 다양한 변수 중에서 2013년의 평균 PM10 농도 및 기온을 지도화한 예시는 Fig. 2와 같다.

II장에서 기술된 바와 같이 Table 1에 해당하는 변수들의 관측소별 원자료를 대기오염 물질의 경우 한 시간 단위로, 기상 변수의 경우 일간 단위로 250개 시군구를 기준으로 추정하였다. 시간별 대기오염 물질 농도와 일간 기상 시군구별 데이터는 CC BY(저작자 표시-비영리) 이용허락 규약을 따르는 것을 전제로 사용 가능하다.

IV. 품질관리

본 데이터의 생성, 유지 및 관리에 관한 모든 사항은 Github로 부여받은 DOI:10.5281/zendo.3685164를 통해 추적 관리된다. 본 데이터의 다운로드, 변경, 사용, 재배포에 대한 제한은 없으나, 원자료의 출처는 위 DOI에 명시된 바와 같이 표기하여야 한다. 본 데이터의 저작권은 Creative Commons의 Attribution 4.0 International(CC BY 4.0) license의 정의를 따른다.

현재 배포된 데이터의 경우 대기과학 및 해당 분야에서 오랫동안 사용되었고, 경험적으로 증명된 방법론이며, 표준적으로 사용되는 오차허용범위(double precision 기준으로 2e-06)에서 모든 추정이 이루어졌다. 배포된 자료의 신뢰성은 다른 자료와의 비교를 통해 확인하였으며, 비교 자료는 Yeo and Kim(2019), Lee and Hong(2019), 환경부의 2018년 대기환경 연보, 그리고 세계보건기구의 Ambient Outdoor Air Quality Database (WHO, 2020)이다.

본 데이터의 품질관리는 세 가지 형태로 진행된다. 첫째, 현재 이 시간에도 원자료의 축적이 계속적으로 이루어지고 있으므로 2018년 이후의 자료에 대해서도 동일한 방법을 적용 데이터를 지속적으로 업데이트할 예정이다. 둘째, 측정방법 및 기술의 발전에 따라 원자료 및 배포 데이터의 품질을 향상할 수 있는 다양한 방법이 존재할 것으로 기대된다. 이들 방법을 적극적으로 활용하여 기존 자료에 대한 업데이트를 진행하며, 변경 및 추가 자료에 대해서는 github의 레코드 추적을 통해 사용자들에게 공개할 것이다. 마지막으로 최종 사용자들에 의해 오류가 보고되면 디버그하고, 요청된 변수들에 대해 추가작업을 진행할 것이다. 추후 개인연구자에 의한 적시 관리에 문제가 발생하면 데이터 품질관리를 위해 관련 기관과의 논의가 필요할 수 있다.

본 데이터는 Yun and Kim(2020b)에서 사용된 바와 같이 일반적인 사회과학적 분석 자료로는 무리가 없는 것으로 판단된다. 하지만, 데이터 구축에 사용된 방법론은 대기과학 및 관련 분야에서 사용하는 가장 단순한 모형이다. 따라서, 대기순환의 복잡성을 반영한 해당 분야의 정교한 모형이 제공할 수 있는 수준의 정확성 및 신뢰성을 보장하지 않는다. 또한 본 자료에는 저자들의 검토과정에서 인지하지 못한 오류나 이상치를 내포할 가능성이 있다. 끝으로 높은 수준의 정확도를 요구하는 경우 혹은 자연과학 및 공학적 측면에서의 분석시에는 데이터 사용에 주의가 요구된다.

적 요

대기오염의 사회경제적 효과에 대한 연구에는 측정된 대기오염 물질, 기상 자료, 그리고 사회경제적 데이터의 병합이 필요하다. 이들 자료들의 시간적・공간적 범위와 단위가 상이하기 때문에 분석에 필요한 데이터 가공에 많은 시간과 노력이 요구된다. 본 데이터의 구축은 사회과학 분야에서 널리 사용되는 대표적인 대기오염 및 기상 변수를 시군구 단위로 제공하는 것을 목표로 한다. 2020년 8월 기준 배포 버전 데이터의 시간적 범위는 2001년부터 2018년이며, 공간적 범위는 250개 시군구로서 패널 형태의 자료를 제공한다. 본 데이터의 기상 변수들은 대기오염 관련 분석뿐만 아니라 다양한 사회과학의 연구에서 사용할 수 있는 주요 변수들을 포함하고 있다.

REFERENCES

AirKorea, 2020: Korea Environment Corporation https://www.airkorea.or.kr/

Beguería, S., S. M. Vicente-Serrano, M. Tomás-Burguera, and M. Maneta, 2016: Bias in the variance of gridded data sets leads to misleading conclusions about changes in climate variability. *International Journal of Climatology* **36**(9), 3413-3422.

CC BY 4.0, Creative Commons Attribution 4.0 International license, https://creativecommons.org/licenses/by/4.0/

Cressie, N., 2015: *Statistics for Spatial Data*. John Wiley & Sons.

Deschênes, O., and M. Greenstone, 2007: The economic impacts of climate change: Evidence from agricultural output and random fluctuations in weather. *American Economic Review* **97**(1), 354-385.

Geoservice, 2020: Geoservice Corporation http://www.gisdeveloper.co.kr/

Hsiang, S., 2016: Climate econometrics. *Annual Review of Resource Economics* **8**(1), 43-75.

KMA (Korea Meteorological Administration), 2020: Climate Information Portal. https://data.kma.go.kr/

Lee, C. J., and M. S. Hong, 2019: Spatiotemporal Variations of Fine Particulates in and around the Korean Peninsula. *Journal of Korean Society for Atmospheric Environment* **35**(6), 675-682.

Schlenker, W., and M. Roberts, 2009: Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences* **106**(37), 15594-15598.

WHO (World Health Organization), 2020: Ambient Outdoor Air Quality Database https://www.who.int/airpollution/data/cities/en/

Yeo, M. J., and Y. P. Kim, 2019: Trends of the PM10 concentrations and high PM10 concentration cases in Korea. *Journal of Korean Society for Atmospheric Environment* **35**(2), 249-264.

Yun, S. D., and B. M. Gramig, 2019: Agro-Climatic data by county: A spatially and temporally consistent U.S. dataset for agricultural yields, weather and soils. *Data* **4**(2), 66pp.

Yun, S. D., and S. G. Kim, 2020a: Air pollution and weather data by Si-Gun-Gu in South Korea (2001 - 2018). https://github.com/ysd2004/AirKorea, DOI:10.5281/zendo.3685164.

Yun, S. D., and S. G. Kim, 2020b: Mismatch of perception and data: Air pollution, medical expenses, and consumption in South Korea. *Environmental and Resource Economics Review* **29**(2), 113-144.