미세먼지 설명변수 전처리

대기오염물질 배출량 및 중국데이터를 중심으로

2017.07.27

김진형

대기오염물질 배출량 데이터

1. SIGUNGU 매칭했을 때 값이 없는 지역

SIDO	SIGUNGU	해결방안
경기	양주군	2003년 양주시로 승격
경기	여주군	2013년 여주시로 승격
경기	안산시	2002년 단원구, 상록구로 나누어짐
경기	용인시	2005년 수지구, 기흥구, 처인구로 나누어짐
경기	포천군	2003년 포천시로 승격
충북	청원군	청주시와 합해지면 됨
충남	연기군	세종시와 합해지면 됨
충남	당진군	2012년 당진시로 승격
충남	천안시	2008년 천안시 동남구, 서북구로 나누어짐
경남	마산시	2010년 창원시 마산회원구, 마산합포구로
경남	진해시	2010년 창원시 진해구가 됨
경남	창원시	2010년 창원시 의창구, 성산구로 나누어짐
제주	북제주군	제주시와 합해지면 됨
제주	남제주군	서귀포시와 합해지면 됨

3. SIGUNGU2 매칭했을 때 값이 없는 지역

SIDO	SIGUNGU	GUNGU2	해결방안
경기	부천시	원미구	부천시로 합해지면 됨
경기	부천시	소사구	부천시로 합해지면 됨
경기	부천시	오정구	부천시로 합해지면 됨
경기	고양시	일산구	2005년 일산동구와 일산서구로 나누어짐
경기	수원시	팔달구	2003년 팔달구와 영통구로 나누어짐

3. 기타

SIDO	SIGUNGU	GUNGU2	해결방안
충북	청주시	상당구	청주시로 합해지면 됨
충북	청주시	흥덕구	청주시로 합해지면 됨
충북	청주시		2014년 상당구, 흥덕구, 청원구, 서원구로

승격: 이름 변경
구가 나누어진 경우: 행 추가 후, 나누어진 수만큼 값을 나눔
구가 합쳐진 경우: 행 aggregation (sum)

대기오염물질 배출량 데이터

연도	시군구	대분류	중분류	소분류	CO	NO_x	SO_x	TSP,	PM_{10}	VOC	NH_3	$PM_{2.5}$

SCC	배출원 대분류
01	에너지산업 연소
02	비산업 연소
03	제조업 연소
04	생산공정
05	에너지수송 및 저장
06	유기용제 사용
07	도로이동오염원
08	비도로이동오염원
09	폐기물처리
10	자연오염원
11	농업
-	-
-	-

SCC	배출원 대분류
01	에너지산업 연소
02	비산업 연소
03	제조업 연소
04	생산공정
05	에너지수송 및 저장
06	유기용제 사용
07	도로이동오염원
08	비도로이동오염원
09	폐기물처리
10	농업
11	기타 면오염원
12	비산먼지
13	생물성 연소(2011년)

자료: 2013 국가 대기오염물질 배출량

- 연도, 시군구, 대분류로 aggregation
- 9개의 대분류별 대기오염물질 7개를 변수로 만들어 9*7=63개의 변수로 만듦

베이징 PM2.5(2008-2016)

- 월단위로 평균
- 베이징으로부터의 거리로 min-max 표준화
  x – max / min - max

Dummy variable

Decision tree의 경우 dummy variable 없이 범주변수를 처리할 수 있음
연도, 월을 범주변수로 처리하려고 함

공간 정보

미세먼지를 추정함에 있어 공간적 변수를 쓰는 실험을 아직 못찾음
향후 다른 연구를 찾아보겠음

Thank you