2. 세부과제별 연구진행상황

3장. 환경 빅데이터 연구

4. 미세먼지 오염도-발생요인 패턴분석 (김진형)

미세먼지 오염도-발생요인 패턴분석

- 연구 필요성 및 연구 목표
  - 대기질에 대한 국민의 관심이 높아졌으나, 미세먼지에 대한 정책은 미흡하고 수도권에 집중되어 있음
  - 미세먼지 발생요인에 정량적 분석을 통해 중요 변수를 선택하고 정책적으로 기여하고자 함

- 연구내용
  - 미세먼지(PM_{10})에 영향을 미친다고 알려진 변수들에 의사결정나무 분석을 적용
  - 종속변수: 2001년~2016년 9월까지 189개월 동안 측정된 미세먼지(PM_{10}) 데이터
  - 설명변수: 기상기후 데이터, 대기오염물질 배출량 데이터, 황사 및 중국 미세먼지(PM_{10}) 데이터, 인구밀도 데이터

- 연구방법
  - 데이터 수집 및 전처리
  - 예측 변수 기술 통계 작성
  - 의사결정나무(랜덤포레스트, bagging, boosting) 분석을 통한 변수 선택
  - 선택 변수의 반사실적 분석을 통한 변수 평가
  - 정책적 제언

활용 데이터 및 전처리 요약

데이터명	시간 해상도	공간 해상도
대기오염물질 농도	1시간	측정소
기상기후 데이터	일, 월, 년 단위	측정소
대기오염물질 배출량	연 단위	시군구
인구밀도	월, 연 단위	시군구
중국 대기질	1시간	측정소
	일 단위	주요 도시

월 단위 데이터로 변환

9개 대분류별 대기오염물질(PM_{10}, NO_x, SO_x, CO, O_3, $PM_{2.5}$, VOC, NH_3) 배출량을 변수로 만듦

인구와 면적을 이용하여 계산 후 사용

베이징, 상해로부터 각 시군구 중심까지의 거리를 이용하여 min-max 표준화함

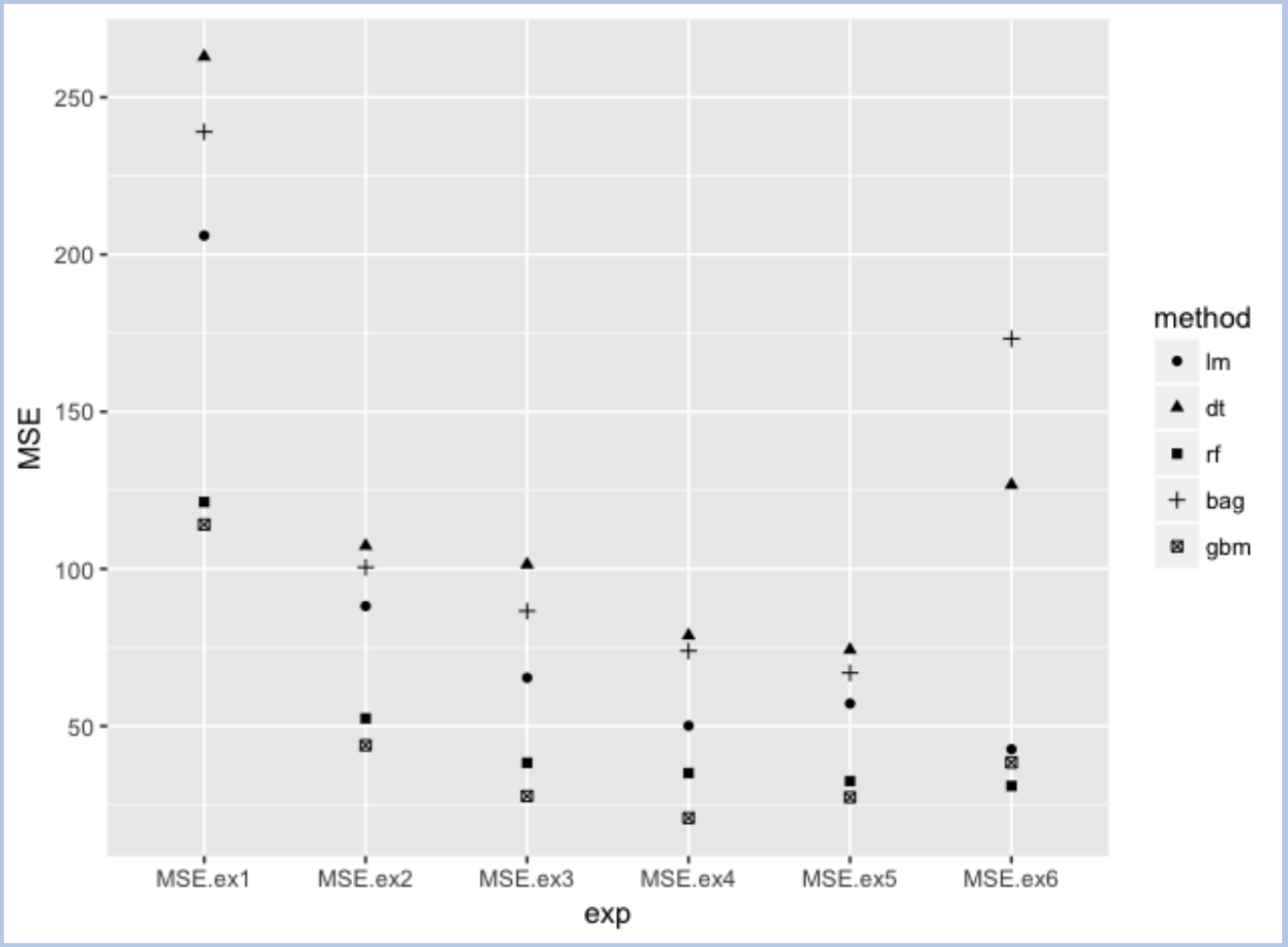
해당 시군구에 매칭

실험설계

- 데이터 별 구축기간 차이로 인해 실험 기간을 달리하여 모델 구축
- 대기오염물질 배출량의 경우 2007년 대분류 변경 이전과 이후로 나누어 실험
  - 실험1의 경우 대분류 변경 이전, 실험2~6의 경우 대분류 변경 이후의 기간을 대상으로 실험
- 중국 대기질 데이터의 경우 구축 시기에 따라 기간을 나누어 실험
  - 실험 1,2의 경우 중국 대기질 데이터를 포함하지 않음
  - 실험 3~6은 베이징 PM2.5 데이터를 포함함
  - 실험 4~6은 상하이 PM2.5 데이터를 포함함
  - 실험 5의 경우, 베이징과 텐진 AQI 데이터를 포함함
- 각각의 기간에 대해 4가지 의사결정나무 분석과 baseline으로써 선형회귀분석을 적용함

실험	실험기간	데이터 설명
실험 1	2001 ~ 2006	대기오염물질 배출량 데이터의 대분류가 변경되기 이전 기간으로 중국 대기질 데이터는 포함하지 않고 실험함
실험 2	2007 ~ 2008	대기오염물질 배출량 데이터의 대분류가 변경된 이후로 중국 대기질 데이터는 포함하지 않고 실험함
실험 3	2009 ~ 2011	대기오염물질 배출량 데이터와 중국 대기질 데이터(베이징)를 포함하고 실험함
실험 4	2012 ~ 2013	대기오염물질 배출량 데이터와 중국 대기질 데이터(베이징, 상하이)를 포함함
실험 5	2014 ~ 2016	대기오염물질 배출량 데이터가 구축되지 않은 기간으로 중국 대기질 데이터(베이징, 상하이, 텐진)를 포함함
실험 6	2007 ~ 2013	대기오염물질 배출량 데이터의 대분류가 변경된 이후로 중국 대기질 데이터(베이징, 상하이)를 포함하고 실험함

모델별 정확도



- 방법론 별 모델 정확도 비교
  - 실험 6을 제외한 실험에서 boosting 모델의 정확도가 가장 높고, 모든 실험에서 random forest 모델이 선형회귀보다 높은 정확도를 가짐
  - Boosting 모델은 실험 1~6의 선형회귀에 대하여 평균적으로 오차를 45.5% 감소시켰으며, random forest는 37.2% 감소시킴
  - 의사결정나무 및 bagging 모델의 경우 선형회귀보다 정확도가 떨어지므로 예측에 사용하는 것은 부적합함
- 실험별 모델 정확도 비교
  - 중국 대기질 데이터를 포함하지 않은 실험 1,2와 비교하였을 때 실험 3~6(의사결정나무와 bagging 결과 제외)의 정확도가 높은 것으로 보아 중국 대기질 데이터를 이용하는 것이 모델 정확도를 높이는 것으로 판단됨

의사결정나무 분석

- 의사결정나무 분석 결과 분기점에 나타난 변수들을 분석하였을 때 다음 표와 같은 결과가 나타남(보고서 p.326~330 참조)
  - MONTH(월)의 경우 모든 의사결정나무에서 분기점으로 나타남으로써 우리나라의 미세먼지 농도의 월별변화가 두드러짐을 보여줌. 또한, MONTH(월)를 예측에 사용했을 때 모델의 정확도가 올라갈 것이라 추정할 수 있음
  - YD_FREQ(황사관측일수)와 함께 BEIJING_PM2.5_STD(베이징PM2.5표준값), SHANGHAI_PM2.5(상하이PM2.5)가 분기점에서 나타나는 것으로 보아 중국 대기질의 영향이 크다고 추정할 수 있음

변수명	실험 1	실험 2	실험 3	실험 4	실험 5	실험 6	SUM
MONTH(월)	1	1	1	1	1	1	6
YD_FREQ(황사관측일수)	1	1	1	0	1	1	5
Y_COORD(위도)	0	1	1	0	0	1	3
MEAN_WATER_PRES(평균수증기압)	0	1	1	0	0	1	3
MIN_SEA_PRES(최소해면기압)	0	1	1	1	0	0	3
BEIJING_PM2.5_STD(베이징PM2.5표준값)	0	0	1	0	1	0	2
SHANGHAI_PM2.5(상하이PM2.5)	0	0	0	1	1	0	2
YEAR(년)	0	1	0	0	0	1	2
MAX_SEA_PRES(최고해면기압)	0	1	0	0	0	1	2
SUM_PRECI(월합강수량)	0	0	1	0	1	0	2

변수중요도

변수명	실험 1_dt	실험 1_rf	…	실험 6_bo	SUM
MONTH	1	1		1	24
MEAN_SEA_PRES	1	0		1	23
Y_COORD	1	1		1	23
SUM_SUN	1	1		1	22
NO2	1	1		1	21
MIN_SEA_PRES	0	0		1	21
SUM_PRECI	0	1		1	21
O3	1	1		1	20
YD_FREQ	1	1		1	20
MAX_SEA_PRES	0	0	…	1	19
X_COORD	0	1		1	19
SO2	0	1		1	17
MAX_WATER_PRES	1	0		1	17
PERC_SUN	1	1		0	17
MAX_TEMP	0	0		1	15
BEIJIING_PM2.5_STD	0	0		1	15
MEAN_TEMP	1	0		0	14
MEAN_WIND_SPED	1	0		1	14
MEAN_MIN_GRA_TEMP	1	0		0	14
⋮					

- 24개의 모델을 통해 도출된 변수 중요도를 통해 중요 변수를 선택할 수 있음(보고서 p.334~ 참조)
  - 모든 모델에서 MONTH(월)가 중요 변수로 선택되었으며, 이는 앞선 실험 결과와 일치하는 결과임
  - X_COORD(경도), Y_COORD(위도)도 중요한 변수인데 지역별 미세먼지의 차이가 반영된 결과임
  - 실험 4를 제외한 5개의 모델에서 YD_FREQ(황사관측일수)가 중요한 변수로 선택되고, 실험 3~6에서 BEIJING_PM2.5(베이징PM2.5), BEIJING_PM2.5_STD(베이징PM2.5표준값), SHANGHAI_PM2.5(상해PM2.5), SHANGHAI_PM2.5_STD(상해PM2.5표준값), BEIJING_AQI(베이징대기질지수), TIANJIN_AQI(텐진대기질지수)가 중요변수로 선정됨으로써 중국 대기질의 영향을 다시 확인함
  - NO2(이산화질소농도), SO2(이산화황농도), CO(일산화탄소농도), O3(오존농도)와 같은 대기오염물질 변수들이 중요 변수로 선택된 것으로 보아 우리나라의 미세먼지 중 2차 미세먼지의 비중이 높을 것이라 예측할 수 있음
  - 기상기후 요인 중에서는 MEAN_SEA_PRES(평균해면기압), MAX_SEA_PRES(최고해면기압), MIN_SEA_PRES(최저해면기압), SUM_PRECI(월합강수량), SUM_SUN(일조시간합), MEAN_WIND_SPED(평균풍속)이 중요 변수로 선택됨. 이 중 SUM_SUN(일조시간합)은 2차 미세먼지 생성에 기여하는 요인으로써 이의 중요성을 다시 확인함

반사실적 실험 결과

컬럼명	30% 감소	10% 감소	10% 증가	30% 증가
SO2(이산화황농도)	-1.922	-0.642	0.630	1.986
CO(일산화탄소농도)	0.131	0.030	-0.003	0.071
O3(오존농도)	0.634	0.118	-0.043	-0.065
NO2(이산화질소농도)	-1.528	-0.485	0.431	1.165
MEAN_PRES(평균기압)	0.941	0.941	-2.602	-2.605
MEAN_SEA_PRES(평균해면기압)	3.225	3.225	-0.994	-0.994
MAX_SEA_PRES(최고해면기압)	4.557	4.557	1.818	1.818
MIN_SEA_PRES(최저해면기압)	3.045	3.045	-2.257	-2.257
SUM_PRECI(월합강수량)	0.466	0.136	-0.123	-0.327
MEAN_WIND_SPED(평균풍속)	0.965	0.285	-0.221	-0.559
SUM_SUN(일조시간합)	-1.054	-0.448	0.501	1.459
BEIJING_PM2.5(베이징PM2.5)	1.483	0.328	0.093	0.703
BEIJIING_PM2.5_STD(베이징PM2.5표준값)	-0.498	-0.157	0.180	0.509
POP_DEN(인구밀도)	-0.038	-0.004	-0.036	0.001
YD_FREQ(황사관측일수)	-0.350	0.003	-0.001	0.435
SHANGHAI_PM2.5(상하이PM2.5)	-3.451	-0.895	0.759	1.710
SHANGHAI_PM2.5_STD(상하이PM2.5표준값)	-0.450	-0.124	0.089	0.239
BEIJING_AQI(베이징대기질지수)	-3.350	-0.915	0.759	1.302
TIANJIN_AQI(텐진대기질지수)	-1.383	-0.341	0.419	1.089

- 가장 높은 예측 정확도를 보인 4개의 모델인 실험 3~5에 Boosting 기법을 적용한 모델, 실험 6에 Random forest를 적용한 모델을 대상으로 반사실적 실험을 수행함
  - 네 모델에서 중요한 변수를 선정하고, 테스트 데이터의 각 변수가 10%와 30%만큼 증가 혹은 감소하였을 때, 모델에서 예측한 PM10 농도의 변화를 살펴봄
  - 미세먼지를 감소시키는 데 가장 큰 기여를 하는 변수는 순서대로 MAX_SEA_PRES(최고해면기압), SHANGHAI_PM2.5(상하이PM2.5), BEIJING_AQI(베이징대기질지수), MEAN_SEA_PRES(최저해면기압)이며, 각각을 30% 증가 혹은 감소시켰을 때, PM10의 농도가 4.56%, 3.45%, 3.35%, 3.23% 감소
  - 미세먼지 생성에 기여하는 대기오염물질 SO2(이산화황농도)와 NO2(이산화질소농도)의 경우 30% 감소시켰을 경우, 각각 1.92%, 1.53%만큼 미세먼지 농도를 감소시킴. 또한, 중국의 공업지역인 텐진시의 대기질 지수를 30% 개선시켰을 때, 1.38%만큼 미세먼지 농도를 감소시킴

결론 및 정책적 제언

1. 중국 대기질 영향
   - 중국의 대기질 데이터를 사용하였을 때, 모델의 정확도가 높아짐
   - 베이징, 상하이, 텐진 시의 대기질 데이터는 변수 중요도 측면에서도 상위에 위치
   - 황사관측일수는 대부분의 모델의 변수 중요도 측정에서 상위에 위치하여 설명력이 높음
   - 월별 변화가 두드러지는 우리나라의 미세먼지 농도 특성 상 황사와 관련한 변수는 예측의 정확도를 크게 높일 수 있을 것으로 예상

2. 국내 NOx, SOx 관리 필요성
   - NOx와 SOx는 높은 변수 중요도 값을 가질 뿐만 아니라 반사실적 실험에서도 PM10 증감에 큰 영향을 미침
   - 일사시간합 또한 중요한 변수인데 빛 에너지가 2차 미세먼지 발생에 많은 영향을 미치므로 NOx와 SOx를 관리해야 함

3. 국내 대기오염물질 발생 패턴 파악 필요
   - 우리나라의 미세먼지 농도는 월별 변화가 매우 큰데 반해 본 연구에서 국내 대기오염물질 발생량은 연 데이터로 월별 변화를 반영하지 못했음
   - 국내 대기오염물질에 대한 세밀한 정책을 위해서는 배출량에 대한 높은 해상도의 데이터 구축과 공개가 필요함