환경 빅데이터 분석 및 서비스 개발

최종자문회의 (2017.10.18)

한국환경정책·평가연구원

강 성 원

1. 연구일반

2. 연구 결과 : 총론

3. 연구 결과 : 2017년 연구성과

4. 결론 및 시사점

5. 향후 계획

1. 연구 일반

개관

구분	내용	
연구성격	일반사업(연구형), 계속사업	
연구기간	2017.1 ~2017.12	
연구진	강성원 연구위원(책임) 장기복 선임연구위원 진대용 부연구위원 홍한움 부연구위원	한국진 전문원 김진형 연구원 김도연 위촉연구원 강선아 위촉연구원 정은혜 위촉연구원 이동현 한국산업기술대 교수(위탁)
자문위원	내부	명수정 연구위원 배현주 연구위원 이명진 부연구위원
	외부	김종률 과장 (환경부 정책총괄과) 우석진 교수 (명지대학교 경제학과) 강희찬 교수 (인천대학교 경제학과) 이성호 박사 (한국개발연구원)
자문일정	착수자문회의: 2017년 3월 30일 중간자문회의: 2017년 6월 29일 최종자문회의: 2017년 10월 18일	

기간, 인력, 예산

- 기간: 2017년 1월 – 2017년 12월
- 인력: 박사급 연구원 5명(1명 원외), 전문원 1명, 연구 보조인력 4명 투입
- 예산: 3억 6백만 원 책정
  - 위탁연구비 4천 만원 책정: '딥러닝을 활용한 환경리스크 예측'
    - 위탁과제 책임자: 한국 산업기술대학교 이동현 교수

목적: 빅데이터 연구방법론 환경정책연구 적용가능성 모색(1)

- 빅데이터 연구방법론(Machine Learning): 기존 연구보다 단기 예측 정확도 제고 및 알려지지 않은 규칙성 발견에 비교우위
  - 예측: 예측오차 축소를 위해 필요한 변인 및 함수관계를 유연하게 선정
    - 기존연구: 변인과 함수관계를 선험적으로 선정하여 예측 오차 발생
    - 단기/개별 관측치 수준 예측에 장점 : 단기/부분적 영향을 미치는 변인을 포괄
  - 규칙성 발견: 변수의 선험적 선택을 지양하여 알려지지 않은 규칙 파악에 유리
    - 기존연구: 규칙성을 파악하고자 하는 변수를 선험적으로 선발하는 경향
- 예측: 단기/소집단 단위 예측 알고리듬 개발 가능
  - 선제적(예측 기반)/ 맞춤형(소지역-개인 특성 반영) 환경정책연구에 활용 가능
- 규칙성 발견: 연구주제/정책대상 발굴 적용 가능
  - 이론적으로 설명되지 않는 규칙성 발굴 : 연구주제
  - 민간 환경관련 문헌 규칙성 발굴 : 연구주제
  - 기존 정책이 포괄하지 못하는 규칙성 발견 : 정책대상

목적: 빅데이터 연구방법론 환경정책연구 적용 가능성 모색(2)

- 빅데이터 연구방법론은 신규 데이터 이용 업데이트 및 유관 분야 연구 적용에 기존 연구보다 유리
  - 변수 간 관계가 자료에 따라 유연하게 설정: 새로운 자료 이용 결과 update 유리
    - 변수 간 관계: 자유도가 높은 함수를 이용하여 자료의 특성을 반영
    - 새로운 함수 설정 없이 기존 연구결과를 신규 자료로 업데이트 가능
  - 자료의 특성이 유사한 연구에 기존 연구에서 개발한 알고리듬 적용 가능
- Update 가능성 활용: 주기적 연구 결과 갱신 가능
  - 관심이 지속되는 동일한 사안에 대한 최신 정보 주기적 갱신
- 유사분야 적용성 활용: 시의성 높은 주제에 대응하는 연구 가능
  - 기본적인 예측을 수행하는 다수의 알고리듬을 개발하여 알고리듬 풀 운영
  - 각 분야의 선험적인 지식보다 자료의 특성을 기준으로 적절한 알고리듬 선택 가능

연구영역: 환경 빅데이터 연구, 인프라, 서비스

환경 빅데이터 연구

- 주제 선정 → 데이터 수집· 가공 → 데이터 분석 전 과정에 빅데이터 분석기법 도입
- 예측 정확도 제고/ 연구주제 발굴/ 주기적 연구/시의성 중심 연구 시행 가능성 모색

환경 빅데이터 인프라 구축

- 연구 자료 및 알고리듬 공개
- 원내외 환경자료 수집-추출 사례 축적 및 공개

원내외 빅데이터 서비스 개발

- 연구성과 활용 원내외 연구정보 서비스 및 공공서비스 개발

연속사업: 3년 단위 연구단계 설정

- 1단계(2017-19): 환경 빅데이터 연구 시작/연구자료 및 분석 알고리듬 공개 시작
- 2단계(2020-22): 환경 빅데이터 분석 플랫폼 설계/빅데이터 활용 공공 서비스 설계
- 3단계(2023-25): 환경 빅데이터 분석 플랫폼 자동화 시도/공공환경 서비스 시범 사업

환경 빅데이터 분석 및 서비스 개발 연차계획

	환경 빅데이터 연구	환경 빅데이터 연구 인프라	원내외 빅데이터 서비스
1기 (2017-19)	• **환경 빅데이터 연구 시행**	• 자료 및 알고리듬 축적/공개	• 원내 연구 및 경영정보 서비스
2기 (2020-22)	• 발신주기 단축	• 빅데이터 연구 과정 자동화 • 환경 빅데이터 분석 플랫폼 설계	• 연구기획 평가 및 준비 서비스 • 공공 서비스 설계
3기 (2023-25)	• 시의성 중심 발신체계 개편	• 환경 빅데이터 분석 플랫폼 지능화 시도	• 공공 서비스 시범 사업

2017년 : 정형화/전산화 자료 분석 중심

- 제 1기 (2017-19년) : 환경 빅데이터 연구 영역 중점
- 2017년 : 전처리 부담이 적은 빅데이터 방법론 적용 가능성 점검
  - 자료의 전처리 부담은 줄이고 다양항 알고리듬 적용 가능성 점검
    - 텍스트 마이닝: 연구주제 파악
    - 수치 자료 분석: 예측 알고리듬 구축
  - 정형화 , 전산화가 되어 있는 수치 및 텍스트 자료 분석에 집중
- 2018년: 전처리가 필요한 자료를 활용하는 빅데이터 연구방법론 점검
  - 비정형, 비전산화 자료: 시청각 자료, 문서로 보관된 텍스트 자료 분석
    - 콘볼루션 신경망(CNN) 등 화상자료 처리에 특화된 알고리듬 적용 가능성 점검
- 2019년: 과거 연구 업데이트 상시화 및 신규 연구 영역 발굴
  - 2017-18년 축적 알고리듬 신규자료 이용 업데이트 → 주기적 발신 작업 시작
  - 텍스트 마이닝 이용 연구주제 파악 상시화

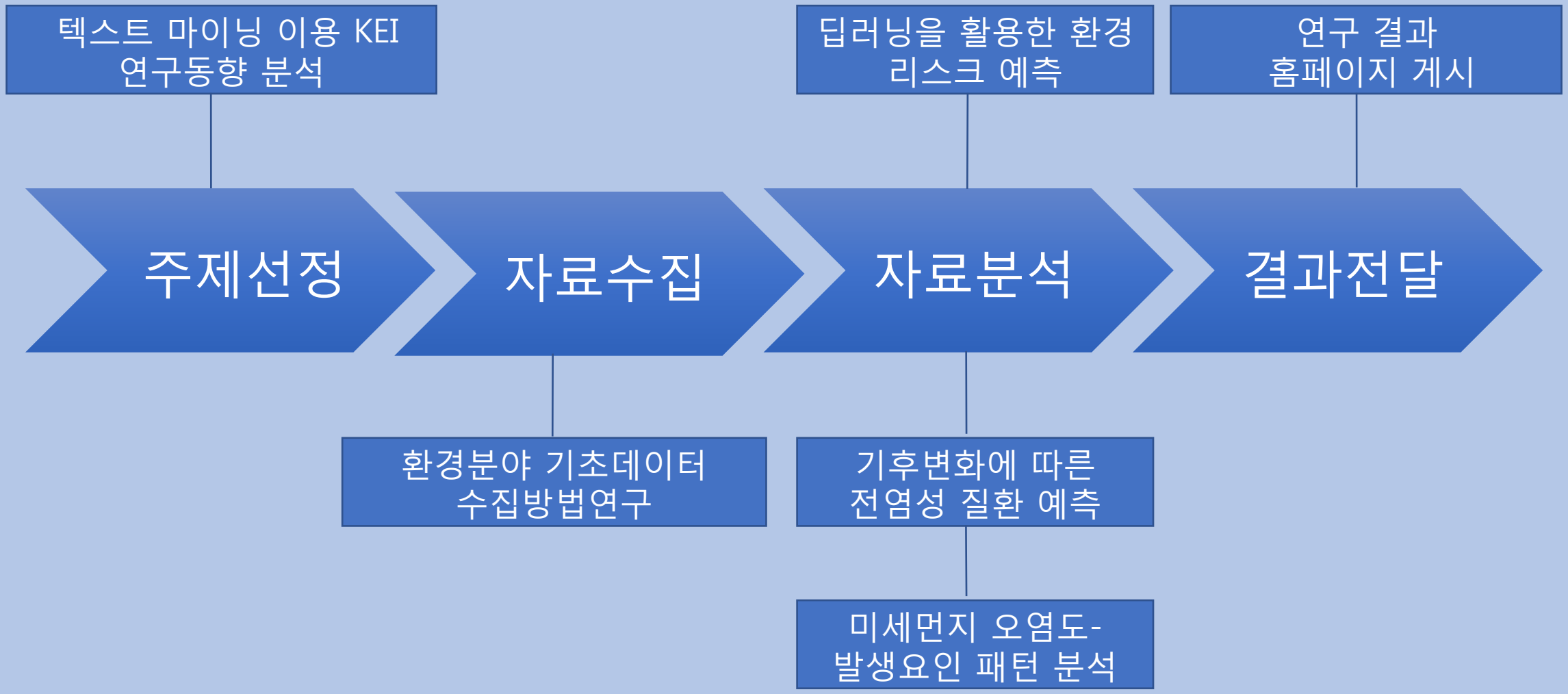
2017-19년 연차계획

	환경 빅데이터 연구	환경 빅데이터 연구 인프라	원내외 빅데이터 서비스
1단계	• **환경 빅데이터 연구 시행**	• **자료 및 알고리듬 축적/공개**	• **원내 연구정보 서비스**
2017	• 환경위험 예측 알고리듬 개발/연구수요 파악: 전산화된 자료 + Deep Learning	• 환경분야 기초데이터 수집방법 • 자료 및 알고리듬 축적/공개	• 연구동향 파악 서비스
2018	• 환경위험 예측 알고리듬 개발/연구수요 파악: 비정형자료 + Deep Learning	• 자료 및 알고리듬 축적/공개 지속 • 환경분야 기초데이터 수집 • 대용량 자료 저장-분석 프로세스 설계	• 연구동향 파악 서비스 원내 • 환경 기초데이터 포털 설계
2019	• 환경위험 예측 알고리듬 개발 : 모든 자료 • 딥러닝 중심연구수요 분석 상시화	• 자료 및 알고리듬 축적/공개 지속 • 환경분야 기초데이터 수집 1단계 완료 • 대용량 자료 저장-분석 프로세스 구축	• 연구동향 파악 서비스 원외 • 환경 기초데이터 포털 원내
2단계	• **발신주기 단축**	• **연구 과정 자동화/플랫폼 설계**	• **연구기획 서비스/공공 서비스 설계**
3단계	• **시의성 중심 발신체계**	• **분석 플랫폼 지능화 시도**	• **공공 서비스 시범 사업**

2017년 세부과제 구성

1. 환경 빅데이터 연구: 환경오염 예측 알고리듬 개발 및 학습 수준 심화
   - 수치 데이터 대상 연구: 예측/패턴분석/원인분석
     - 시계열 자료 예측 : 환경오염 예측 딥러닝 알고리듬 개발 (순환신경망 모형)
       - 오염 예측의 시간-공간 해상도 제고
     - 시공간 자료 패턴 분석 :기후자료-건강보험 자료 패턴 분석 (심층신경망 모형)
     - 환경오염 원인 규명: 미세먼지 발생 요인과 오염도 간 관계 규명 (의사결정나무 모형)
   - 텍스트 데이터 대상 연구: 연구동향 파악
     - 자연언어 분석기법을 활용한 KEI연구보고서 및 인터넷 환경뉴스 분석
       - 토픽 분석: 연구 주제 구성의 시간적 흐름 비교 분석
       - 키워드 연관어 및 네트워크 분석 : 연구 키워드간 관계의 시간적 흐름 비교분석
       - 문장 내 단어간 연관성 분석: 세부주제의 강조점 비교 분석
2. 환경 빅데이터 인프라 구축: 원내외 환경관련 자료 수집-추출 사례 축적
   - 산재된 환경 관련 자료를 수집–추출하는 사례를 축적하여 오픈 소스로 공개

세부과제 구성


텍스트 마이닝 이용 KEI
연구동향 분석
딥러닝을 활용한 환경
리스크 예측
연구 결과
홈페이지 게시
주제선정
자료수집
자료분석
결과전달
환경분야 기초데이터
수집방법연구
기후변화에 따른
전염성 질환 예측
미세먼지 오염도-
발생요인 패턴 분석

중간자문회의 자문의견 반영결과

- 전체 맥락에서 세부과제 선정 근거 불명 : 2017-19년 연차계획 내 선정근거 반영
- 정책적 시사점 보완 필요 : 결론에 정책적 시사점 별도 작성

위 원	제 안 내 용	반 영 내 용 (미반영시 사유)
우석진	- 정책적 활용 및 기여도에서 좀 더 구체적으로 고민해야 할 것으로 판단됨.	- 결론에서 정책적 시사점을 별도로 언급
이성호	- 이동현 박사 연구의 경우, 현재를 빠르게 예측하는 now casting에 역점을 둔다면 실용성을 제고할 수 있을 것으로 사료됨.	- 현재 2시간 예보가 가능한데 2일(48시간) 예보가 가능하도록 모형을 확장할 계획
김종률	- Deep learning을 통한 리스크 예측 개발 중 미세먼지의 경우 서울시 이외 다른 지역 확장 보다는 서울시 분석결과를 심층분석/해석/보완을 통해서 연구의 질을 제고하는 것이 필요하다고 생각함. - 분석방법에 따라 결과가 다른 게 나오는 경우(GAM vs OLS), 이런 원인에 대한 설명과 대안 마련이 필요	- 서울시 분석 결과를 2시간 예보에서 2일 예보로 심화하고 지역 확장은 점진적으로 진행 - GAM은 종속변수와 독립변수의 관계가 비선형일 경우 적용 /OLS 는 종속변수와 독립변수의 관계가 선형일 경우 적용. 추정 결과를 보다 면밀히 검토하여 현실정합성이 높은 결과를 수용할 예정
강희찬	- 전체 맥락에서 특정 인과·상관관계를 연구주제로 선택한 이유에 대한 충분한 설명이 필요함. - 합리적 차원에서 금년 내 완수가 어려운 부분에 대한 합의 및 팀원간 공유 필요. - 딥러닝을 통한 예측 방법론은 미세먼지 분야보다는 수질 등 타 분야에 적용하는 것이 보다 많은 기여도가 있을 것으로 사료됨.	- 서론에 보강 - 7,8,9월 3차에 걸쳐 중간성과를 점검하고 최종결과물을 조정할 예정 - 2년차 이후부터 수질 분석으로 연구대상을 확장할 예정

중간자문회의 자문의견 반영결과

위 원	제 안 내 용	반 영 내 용(미반영시 사유)
명수정	- 연구 내용이 방대하여 하루-이틀 간의 세미나 개최를 추천함. - 세부과제의 내용이 전체 큰 틀에서 갖는 역할에 대한 소개 내용이 보완될 필요. - 일반적으로 자료수집(특히 해외자료)과 분석은 쉽지 않아 단기간 내 반영이 어려우므로, 여건을 고려하여 1차년도 이후 중장기로 반영하는 것을 고려하시기 바람. - 보고서 구성과 배치, 소제목과 부록 등에 대하여 좀 더 고민이 필요함. 장별 구성보다는 부별 구성을 제안. - 중장기적 계획을 구체화할 필요. 현재 연구 중인 내용의 지속 또는 추가 여부에 대한 시기적 계획 필요. - 중국 대기오염 데이터 외 인구와 산업 자료 등 다양한 관련자료 활용을 고려. - 미세먼지 분석의 경우 시군구별 분석이 이루어진다면 해석과 정책적 활용성을 제고할 수 있을 것으로 사료됨. - 보건은 질병 특성을 고려할 필요. (예. 노로 바이러스 관련 질환은 겨울철에 더 자주 발생) - 정책기능 분류에서 '기후'는 '부문'이라기보다 '영향'이므로, 결과 도출 후 재배치 필요. - 정책적 함의에 대한 내용 확충 필요. - 기존 환경 연구자들과의 논의를 통하여 환경 분야 데이터 수요를 파악할 필요.	- 7,8,9월 진행점검 세미나에 반영 - 서론에 연구로드맵을 보다 정교하게 작성 - 서론의 연구로드맵에 반영 - 보고서 구성을 1부-2부로 나누고 수량변수 추정 연구를 2부에 집중 - 서론의 연구로드맵에 반영 - 중간보고 이후에 시간 제약하에서 반영할 수 있는 부분을 반영하고 나머지는 중장기 과제에 포함 - 질병 특성에 대해 소개하는 내용을 포함 - 정책기능 분류 재점검 시 고려 - 보고서 결론에 반영 - 심층 설문조사를 추진할 계획

중간자문회의 자문의견 반영결과

위 원	제 안 내 용	반 영 내 용(미반영시 사유)
배현주	- 연구과정과 결과 검증에 대한 전문가들의 참여와 충분한 의견수렴 과정이 포함되었으면 함. - 세부과제별 연구 성과물 완성도가 상당히 차이가 크므로 최종보고서에 어떻게 정리하여 담을 것인지에 대한 고민이 필요.	- 7,8,9월 3회 중간성과 점검회의 시 자문위원 참석 유도 - 중점적인 과제의 완성도가 더 높은 상황이므로 과제 우선순위에 따라 성과를 소개하는 순서로 구성할 예정
이명진	- 세부과제를 아우르는 큰 흐름에 대한 설명이 서두에서 확충될 필요. - 자료의 공간적 해상도 등 보다 정밀한 자료의 선행 특성 분석이 강화될 필요가 있음. - Decision Tree에 자체 프로그래밍이 보다 필요함. - 중간보고회 이후의 실제 모델링에 자료를 연계하여 분석하는 과정이 중요하며, 특히 결과에 대한 중립적 해석을 보다 강화하는 것이 필요함. - KEI에 국한되지 않고 환경 전 분야로 확대된다면 정책적 기여도가 보다 높을 것으로 생각됨. - 다양한 환경관련 데이터에 대한 전문지식을 가진 연구진이 충원될 필요.	- 서론부문 연구로드맵에 반영 - 하반기 연구 시 검토 - Decision Tree는 다양한 package가 개발되어 있어서 자체 프로그래밍 보다는 package를 이용하는 분석이 더 효율적 (미반영) - 결과 해석 시 반영 - KEI 문헌 중심 세부과제들은 분석 대상 문헌의 범위 및 수량을 점진적으로 확대할 예정 - 상반기에 데이터 분석 경험이 있는 연구진 2명을 이미 충원하여 추가적 연구진 충원은 어려움(미반영)

보고서 목차 및 작업계획

부	장	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
1부: 총론	서론										
	빅데이터 연구방법론 활용방안 (강성원)										
2부. 2017년 연구성과	딥러닝을 활용한 환경리스크 예측 (이동현)									후속	조치
	기후변화에 따른 전염성 질병 예측 (강선아)										
	텍스트 마이닝 이용 KEI 연구동향 분석 (김도연)										
	미세먼지 오염도-발생요인 패턴 분석 (김진형)										
	환경분야 빅데이터 수집 방법론 (한국진)										
3부: 요약 및 시사점	2017년 연구성과 정리										
	정책적 활용방안										

연구 관리

- 월 1회 진도 점검 세미나 진행
  - 프로포절 세미나(4월)
  - 진도점검(5, 7, 8, 9월)

- 기존 연구성과 온라인 공개
  - Homepage: https://keibigdata.github.io/index.html
  - Github repository: https://github.com/keibigdata

Homepage

KEI Bigdata Research Team Blog

Bigdata 연구방법론 활용방안

[Bigdata 연구방법론 활용방안] Proposal (pdf)

Posted by **Sung Won Kang** *on April 13, 2017*

[Bigdata 연구방법론 활용방안] Progress Report (May) (pdf)

Posted by **Sung Won Kang** *on May 24, 2017*

딥러닝을 이용한 기후변화에 따른 전염성 질환 발생 패턴 분석

[전염성 질환 발생 패턴 분석] Proposal (pdf)

Posted by **Suna Kang** *on April 13, 2017*

[전염성 질환 발생 패턴 분석] Progress Report (May) (pdf)

Posted by **Suna Kang** *on May 24, 2017*

keibigdata / sungwonkatto2

Unwatch 1 Star 0 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights Settings

No description, website, or topics provided.

Edit

Add topics

12 commits 1 branch 0 releases 1 contributor

Branch: master New pull request Create new file Upload files Find file Clone or download

katto2 Update README.md Latest commit 5943b03 on 27 Jul

ENVpapers_0725.csv	Add two files for Environmetal Policy Paper analysis	3 months ago
EnvEssary_progress_July.pptx	add ppt for July progress	3 months ago
EnvResearch.csv	first upload	4 months ago
EnvResearch_0619.csv	first upload	4 months ago
EnvResearch_budget.csv	first upload	4 months ago
EnvResearch_summary.Rmd	first upload	4 months ago
EnvResearch_summary.html	first upload	4 months ago
Envpapers_summary.Rmd	Add two files for Environmetal Policy Paper analysis	3 months ago
Envpapers_summary.html	add Envpapers_summary.html reporting Environment policy paper analysis	3 months ago
QRESEARCH.csv	QRESEARCH	4 months ago
README.md	Update README.md	3 months ago

README.md

sungwonkatto2

2016.07.05 현재 등재 파일

- 자료 DB

2. 연구 결과 : 총론

2. 연구 결과 : 총론

2장 빅데이터 연구방법론 활용방안

빅데이터 연구방법론 활용방안

- 환경정책연구 방법론과 빅데이터 분석기법의 특징을 파악하여 적용 방안 도출
  1. 환경정책연구: 환경정책을 유형화 및 유형별 관련연구 특성 파악
  2. 빅데이터 분석방법: 주요 분석기법 정리 및 장점 파악
  3. 환경정책연구 유형별 빅데이터 분석기법 활용방안 도출
- 환경정책 유형별 관련연구 방법론 특징 추출
  - 환경정책 유형화 : 예산서, 환경백서 이용 유형 도출 후 연구문헌 유형별 분류
  - 분석 대상 연구문헌: 2016년 KEI 보고서 + 2013~17 환경정책연구 학술논문 (Google Scholar 검색 결과)
  - 유관 연구 유형별 방법론 분포 파악: 정량적 방법론 비중 파악
  - 정량적 방법론 사용 목적 파악 : 빅데이터 분석 방법 적용 여부 점검

환경정책 유형별 관련연구 특징 추출

- 환경정책 유형화 : 부문 및 기능
  - 부문: 예산서 '관' 항목을 사용하되 '환경일반'을 세분화
    - '관' 항목: '상하수도', '수질', '폐기물', '대기', '자연환경', '환경일반'
    - '상하수도'와 '수질'은 '상하수도-수질'로 통합, '자연환경'에서 환경영향평가를 구분
    - '환경일반': '화학물질', '환경보건', '환경산업(기술, 경제)', '국제협력' '기타'

  - 기능: 환경오염물질 발생단계에 따라 정책기능 분류
    - 억제(Control): 배출원의 환경오염물질 배출을 억제하는 기능
      - 환경규제/환경관련 부담금
    - 처리(Treatment): 이미 배출된 오염물질의 영향을 저감하는 기능
      - 오염물질 처리 설비(환경기초시설) 및 시설(하수종말처리장, 폐기물매립장) 설치/환경오염 피해 보상
    - 대비(Preparation): 배출 이전에 배출량 저감 조치 유도 혹은 배출 정보 제공
      - 환경영향평가/환경오염예보
    - 환경조성(Support): 여타 3가지 기능이 원활하게 수행될 수 있는 여건 조성

환경연구 분포

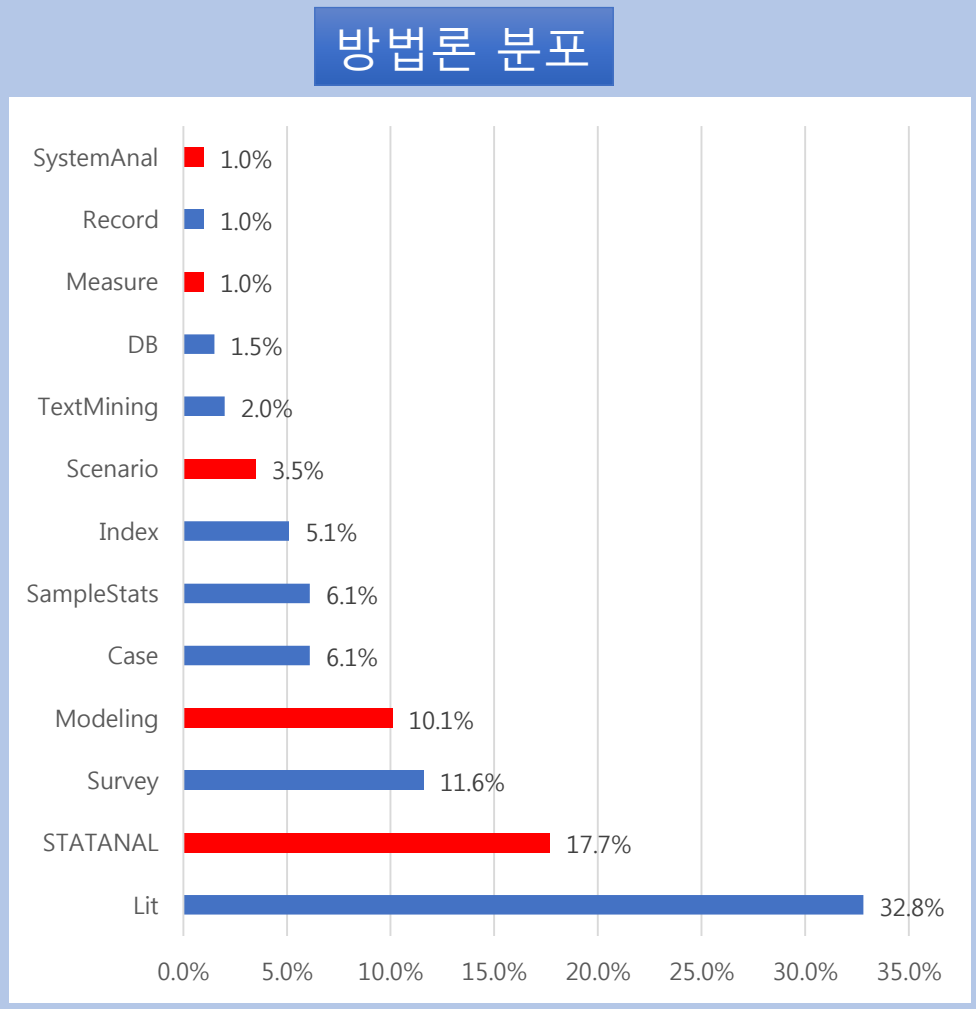
		억제 (규제. 조세)	예방 (예보, 영향평가)	환경조성 (계획.교육,법제도)	처리 (시설.복구)
상하수도,수질		3	3	3	12
폐기물		0	0	2	6
대기(기후)		39	2	11	8
자연환경		7	5	6	12
영향평가		0	8	1	0
환경일반	화학물질	2	0	1	0
	환경보건	0	0	3	1
	환경산업	1	0	6	0
	국제협력	0	3	14	0
	기타	3	5	30	0

환경정책연구 방법론 분류기준

명칭	내용
STATANAL*	통계적 분석 방식을 사용하여 변수간의 관계를 파악하거나 관심변수의 값을 추정하는 방법 - 결정요인분석, CVM, Conjoint, 메타연구 등 중간단계에서 계량경제학적 분석방식을 요하는 방법론 포함 - 클러스터 분석, 주성분 분석 등 전통적인 회귀식을 이용하지 않는 통계적 추론 방법론 포함
Modeling*	연역적 추론에 기반한 모형을 구축하고 이를 이용하여 관심 변수의 값을 구하는 방법 (일반균형, 산업연관분석)
SystemAnal*	사전적 모형 없이 직관적인 인과관계 네트워크 시스템 모형을 구축하여 분석하는 방법
Scenario*	파라미터 값이 상이한 시나리오를 구축하고 관심 변수의 값을 시나리오에 따라 구하는 방법 (경제성, 수익성..)
SampleStats	특정한 방법론 없이 기초자료로부터 표본통계량을 조합하여 논거를 찾아내는 방식의 연구를 의미
Index	기초통계량으로부터 관심대상 현상을 대표하는 지표를 도출하는 방식
DB	DB구축
Lit	주제와 관련된 선행연구 및 사례에 관련된 문헌을 종합하여 정리하고 시사점을 도출하는 방법 - 정량적 연구 중 결과물을 도출하지 않고 방법론을 정리한 연구도 문헌연구로 간주
Survey	설문조사(추가적인 분석 없이 설문조사 결과만 제시한 경우)
SpecialCouncil	전문가들로 구성된 패널의 의견을 종합하는 방법
Record	행사기록
Case	문헌조사 이외의 방법을 사용하여 사례를 조사하는 방법(인터뷰, 실측 등을 포함)

문헌조사가 방법론의 압도적 비중 차지

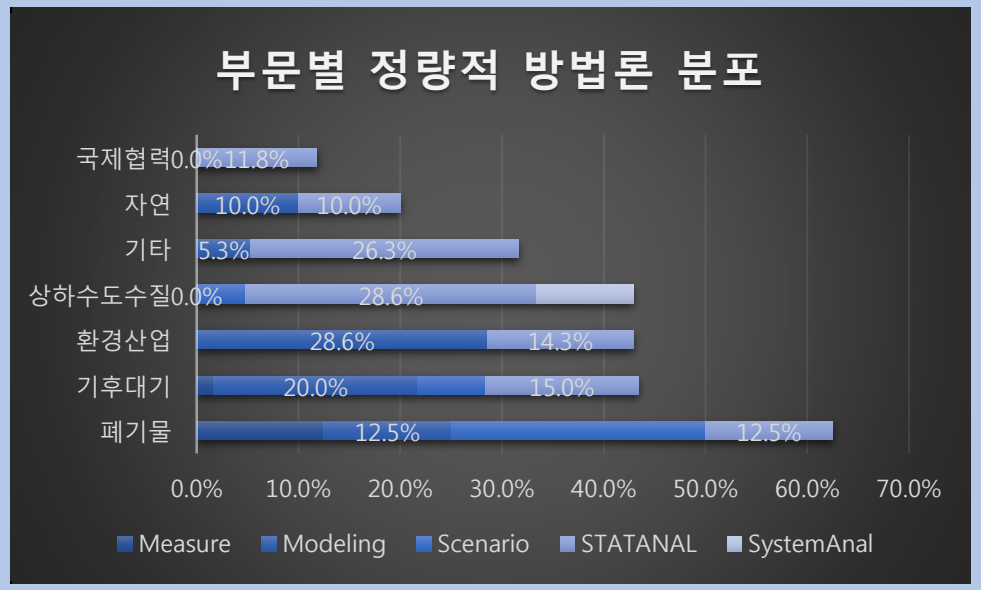
- 문헌연구 비중: 32.6%
- 통계분석 비중: 17.7%
- Machine Learning 과 유사한 '정량적' 연구는 36.4%
  - 통계분석 (17.7%)
  - 모델링 (10.1%)
  - 시나리오 분석 (3.5%)
  - 측정법(Measure 1.0%)
  - 시스템 분석 (1.0%)

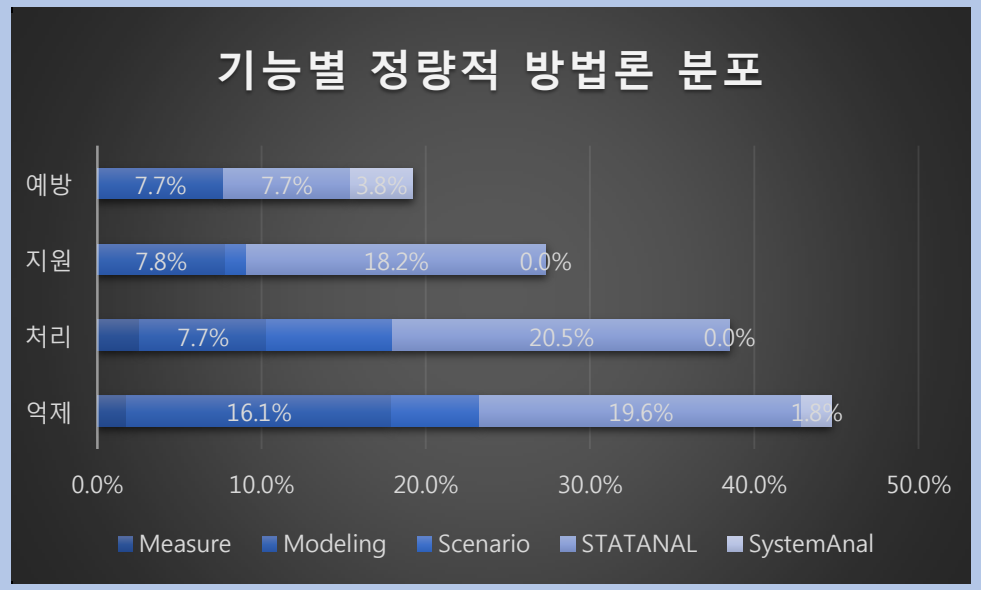


폐기물부문, 억제기능: 정량연구 비중 높음

- 부문: 폐기물(62.5%), 기후대기(43.4.1%), 환경산업(42.9.%).상하수도수질(42.0%) 정량연구 비중 40% 이상
  - 전체 연구 건수가 5건 이하인 부문은 제외
- 기능: 사전적 억제 (44.7%), 사후적 처리(38.5%)의 정량연구 비중이 높음
  - 지원(27.3%) : 제도 관련 연구의 비중이 높아서 문헌연구의 비중이 높음
  - 예방(19.2%) : '환경영향평가' 관련 연구의 비중이 높은데 환경영향평가 관련 연구는 문헌 연구 비중이 높음

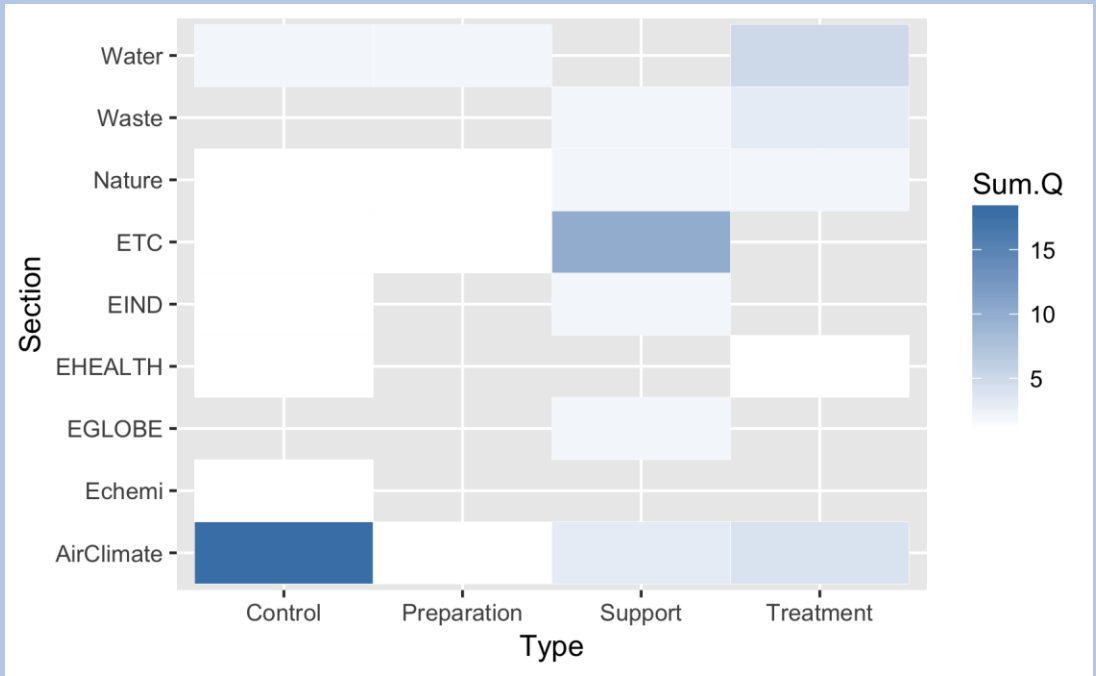
부문별, 기능별 정량적 연구 비중



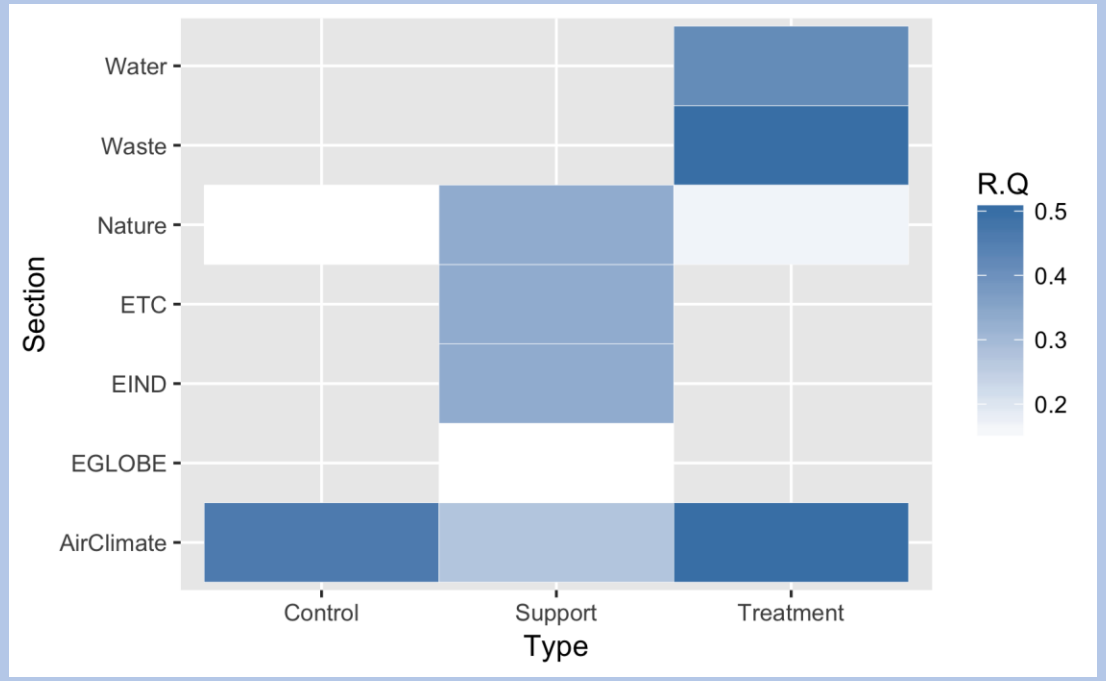


부문-기능 별 정량연구 비중

부문- 기능별 정량 연구 건수



부문-기능별 정량 연구 비중

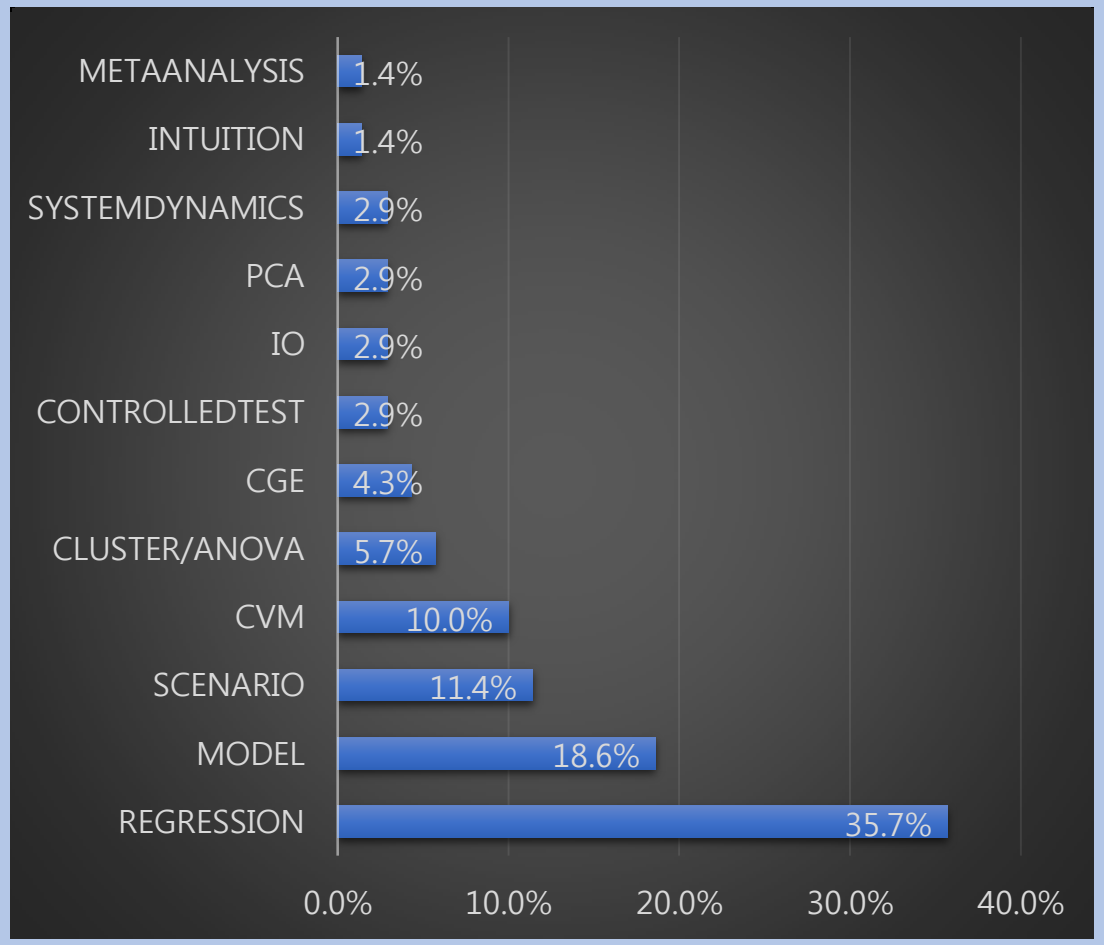


정량연구는 예측 목적으로 주로 수행

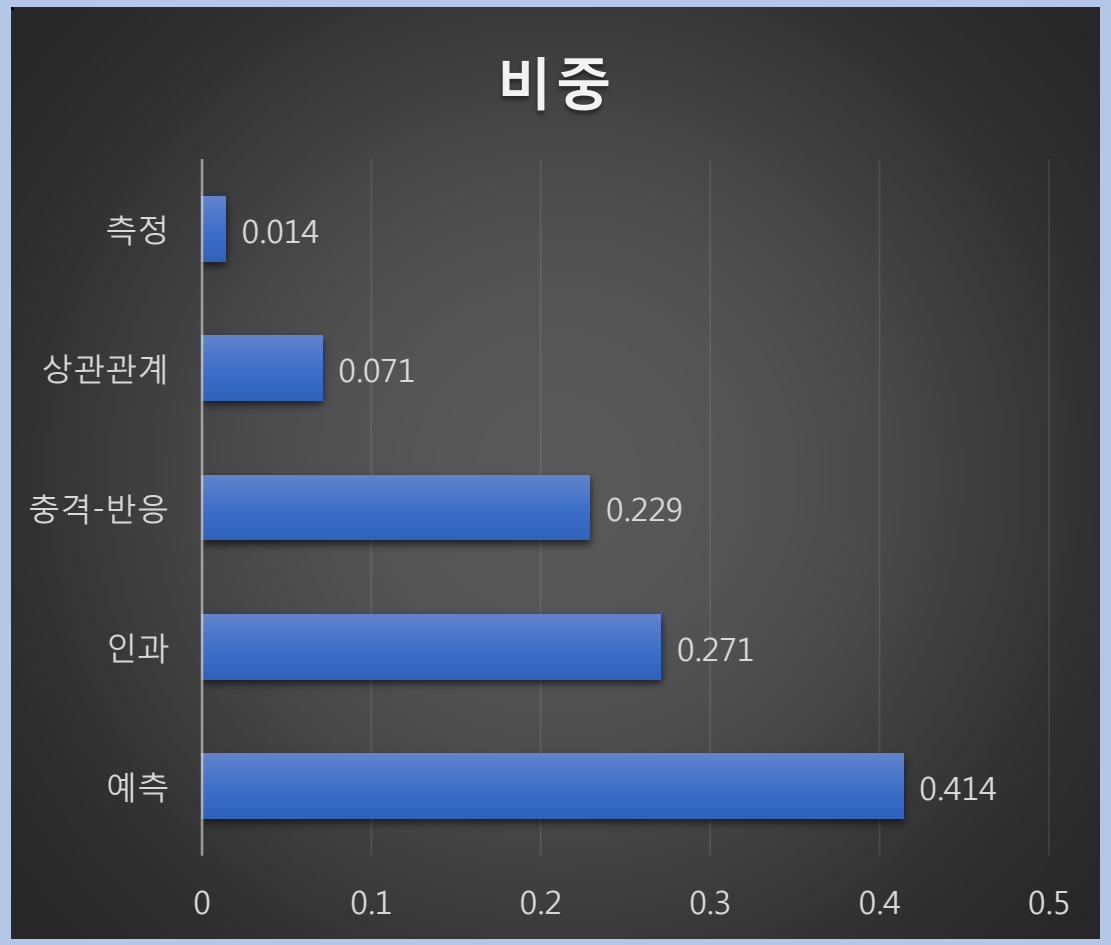
- 정량연구 방법론: 선형회귀(35.7%), 부문별 모형(18.6%), 시나리오 분석 (11.4%) 활용 활발

- 정량연구 목적: 예측 (41.4%) > 인과(27.1%) > 영향-반응 (22.9%)
  - 예측: 관심변수의 미래 값 추정
  - 인과: 특정 변인이 관심변수에 미치는 영향의 존재 여부/방향
    - 양적인 반응에는 관심이 없는 경우
  - 영향-반응: 변인의 변화에 따른 관심변수의 양적인 변화량
    - 인과관계로 인한 관심변수의 양적 변화에 관심이 있는 경우

정량적 연구 방법론 분포

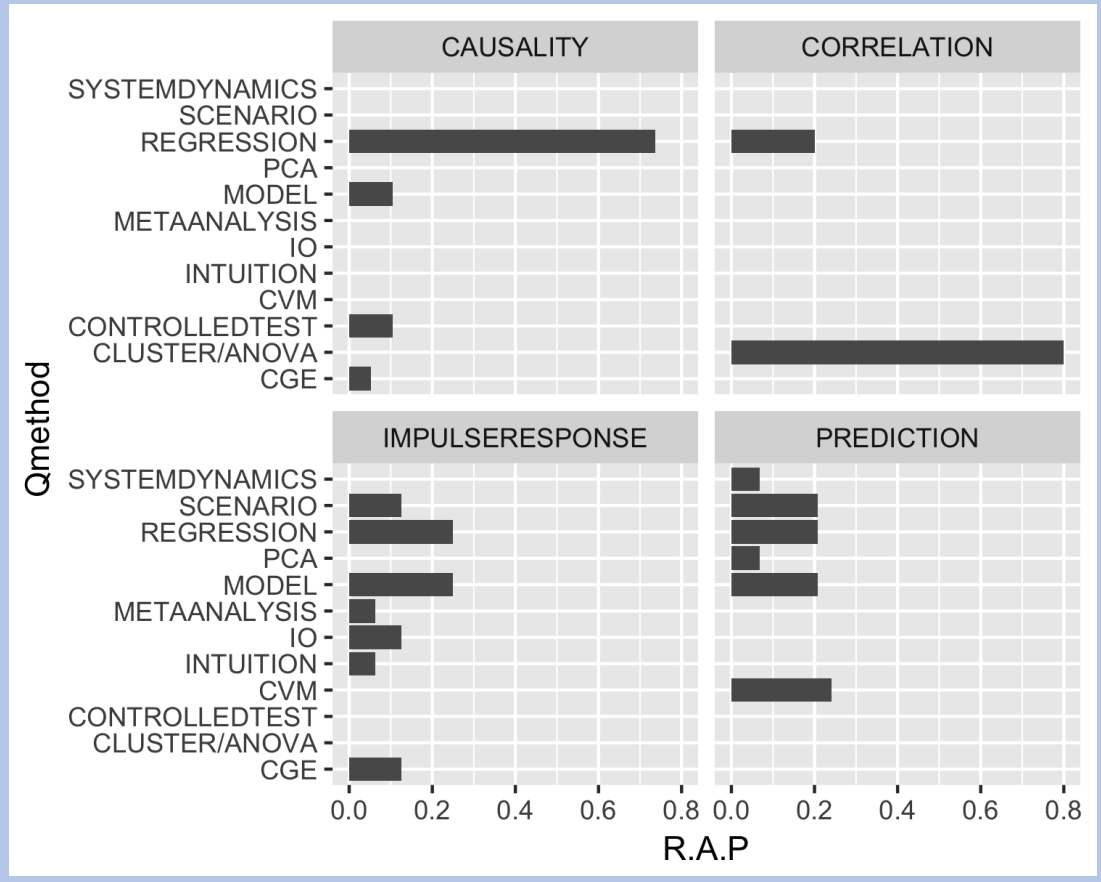


정량적 연구 목적



예측, 영향-반응 추정: 비통계적 접근 빈번

- 인과, 상관관계: 회귀분석/군집-분산 분석
  - 인과분석 : 73.7% 회귀분석
  - 상관관계 : 80% 군집-분산 분석
- 예측, 영향-반응 추정 : 비 통계 방법론 비중 높음
  - 예측: 부문별 모형(20.7%), 시나리오(20.7%)
  - 영향-반응: 부문별 모형(25.0%), 일반균형분석 (12.5%), 시나리오 분석 (12.5%)
- 비통계적 방법론: 변인에 대한 선험적 지식 부족/변인-관심변수 간 통계적 추정 어려움 대응
  - 관심변수-변인 관계가 선험적으로 알려져 있지 않은 경우: 시스템 분석, 시나리오, 메타분석, 직관
  - 관심변수-변인 관계가 통계적 추정이 어려운 경우: 부문별 모형, CGE



빅데이터 분석기법: 경험적 변인선정/관계 파악

- 빅데이터 분석기법 : 기계학습(Machine Learning)
  - 지도학습(Supervised Learning) : 관심의 대상이 되는 종속변수를 다양한 변인을 활용하여 예측하는 알고리듬을 개발
    - 표본을 학습자료, 평가자료로 분할: 학습자료로 관계파악, 평가자료의 예측오차로 학습성과 평가
    - 변인의 선정(정규화): 예측 오차를 최소화하는 변인의 조합을 선정
    - 관심변수-변인 관계(학습): 자유도가 높은 함수의 파라미터를 예측오차를 최소화하도록 선정
    - 회귀분석(연속 종속변수 예측), 분류(이산 종속변수 예측) 등 과제 수행
  - 비지도학습(Unsupervised Learning): 자료의 변수 간 상호관계를 파악하여 패턴을 도출
    - 군집 도출(자료 분류), 연계규칙 학습(규칙성 파악), 차원 축소(관계 설명 변수 도출) 등 과제 수행

- 빅데이터 분석기법의 장점: 단기-개별 관측치 수준 예측에 유리
  - 기존 정량연구: 선험적(이론적)으로 변인을 선별 – 평균적, 장기적 영향이 있는 변인을 선택
  - 빅데이터 분석: 예측에 도움이 되는 모든 변인을 포괄 – 부분적, 단기적 영향이 있는 변인 포괄

기계학습 세부방법론 분류

label	과제	조건부평균/조건부확률		세부방법론
지도학습 (label 존재)	회귀분석	알려진 함수	선형	변인 변형 없음: 선형회귀, 라소회귀, 리지회귀
				변인 차원 축소: 주성분회귀, .부분최소자승
			비선형	GAM, 서포팅 벡터 회귀
		알려지지 않은 함수		의사결정나무 회귀, KNN, 심층신경망
	분류	알려진 함수		분류 특화: 로지스틱 회귀, 나이브 베이즈, LDA, QDA 분류 및 회귀분석: 서포팅 벡터 메커니즘
		알려지지 않은 함수		분류 및 회귀분석: 의사결정나무 분류, kNN, 심층신경망
비지도학습 (label 부재)	군집			클러스터, 평균-최대화 알고리듬
	연관규칙			Aprior, Eclat, FP-growth
	차원축소			주성분분석, 커널주성분분석, 로컬선형임베딩, 통계적 임베딩

주 1. 회귀분석 방법론은 분류 방법론으로 활용 가능
주 2. 기존 방법론을 결합한 앙상블(Ensemble) 방법론 활용 가능

빅데이터 분석방법 활용: 비 통계 방법론 보완

- 지도학습 방법론: 정량연구 중 비 통계 방법론 보완 가능
  - 관심변수-변인 관계 선험적 지식 부족: 정규화 과정을 통해 예측오차를 줄이는 변인을 선정
  - 관심변수-변인 관계 통계적 추정 어려움 : 자유도 높은 함수 파라미터를 예측 오차 축소할 수 있도록 학습
- 지도학습: 비 통계적 연구 방법론 비중이 높은 예측/영향-반응 연구 보완 가능
  - 예측연구 : 정규화 과정을 통해 변인을 선정하고, 학습과정을 통해 관심변수-변인간 관계를 통계적으로 파악
  - 영향- 반응 추정 연구: 반 사실적 실험을 활용하여 변인의 가상적 변화에 대한 관심변수 정량적 반응 측정
- 비지도학습: 상관분석 방법론 다변화 및 연구주제 발굴
  - 클러스터 분석 의존도 완화
  - 기존의 이론으로 설명하지 못하는 잠재적 변인과 관심 변수(환경오염도) 상관관계 파악: 연구주제 제기
- 한계: 단기/개별 관측치 수준의 예측 정확도 제고에 유리하나 인과분석은 어려움
  - 중장기/평균 수준 예측: 부분적인 표본에 영향을 미치는 변인의 영향이 소멸하여 기계학습의 이점이 사라짐
  - 인과분석: 관심변수와 변인 간 명시적(Explicit) 관계 파악이 어려우며, 통계적 유의성 판정이 어려움
    - 축약형(Reduced form). 관계 파악에 그침: 변인이 관심변수에 영향을 미치는 관계는 black box
    - '상관관계는 파악이 가능하지만 인과관계는 파악이 어려움'

3. 연구 결과 : 2017년 연구성과

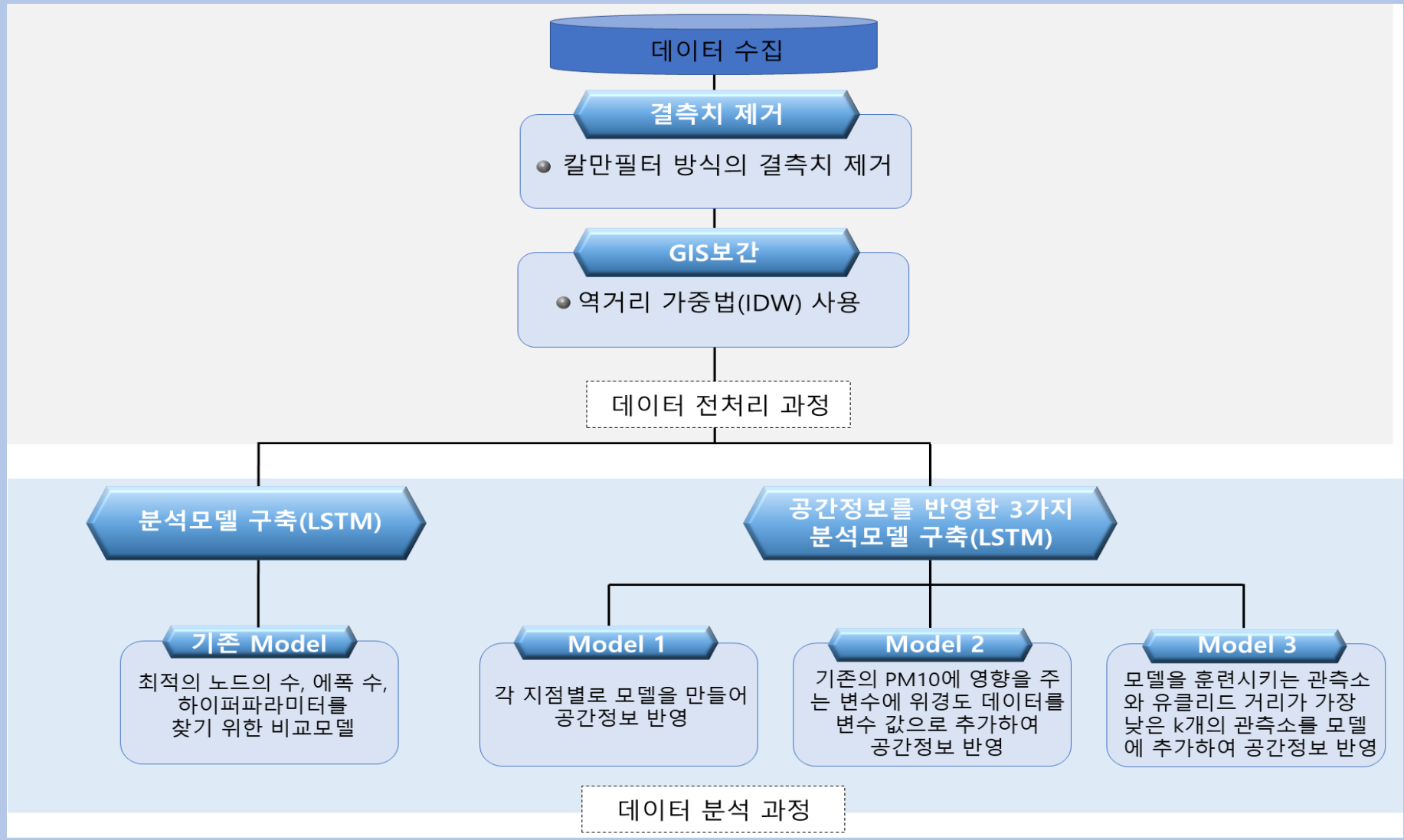
3. 연구결과 : 2017년 연구성과

1. 딥러닝을 활용한 환경 리스크 예측 (이동현)

(1) 딥러닝을 활용한 환경 리스크 예측

- 딥러닝 기술을 활용하여 전국 측정소 단위에서 미세먼지의 시간당 오염도를 예측
  - 종속변수: AirKorea 미세먼지(PM_{10})
    - 2016년 1월 ~ 2016년 12월 서울시 25개 도시/14개 도로변 관측소 측정 데이터
  - 독립변수
    - AirKorea 대기오염물질 (아황산가스, 일산화탄소, 오존, 이산화질소)
    - 기상청 방재기상관측 데이터 (시간당 기온, 풍속, 풍향 등)
  - LSTM, kNN공간순환신경망: ARIMA 평균제곱근오차(RMSE) 10% 개선

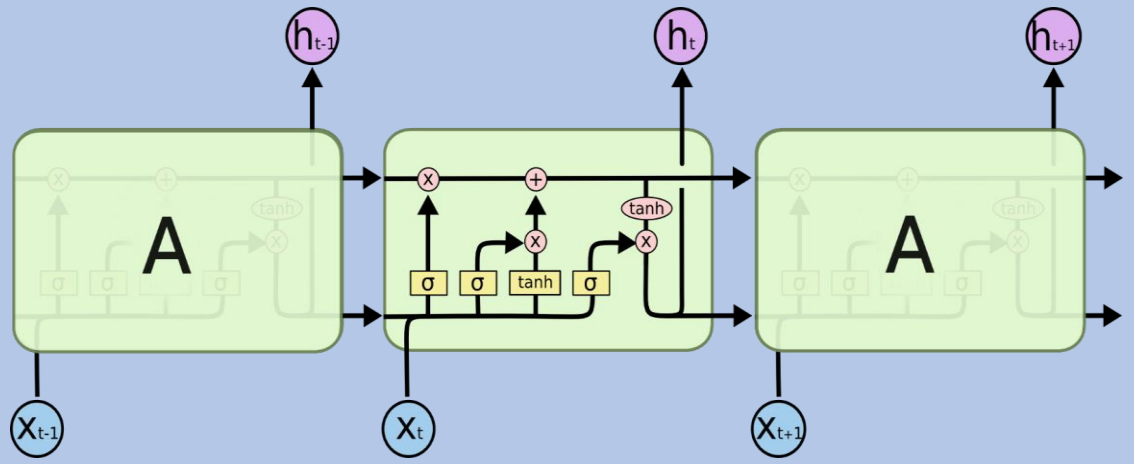
연구 Framework



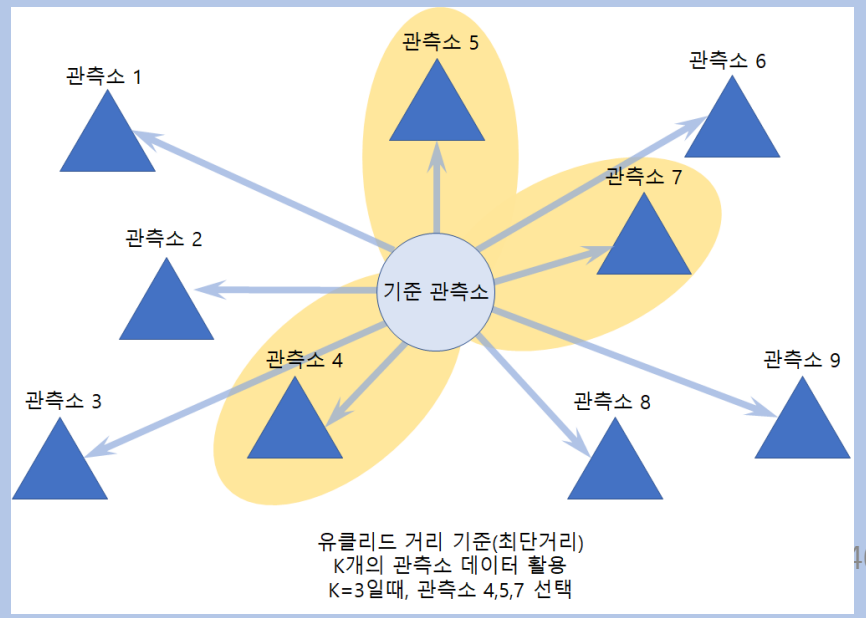
분석방법: 다층순환신경망, kNN공간순환신경망

- LSTM: 해당 측정소 과거 관측치 정보 반영
- kNN 공간순환신경망: 과거 관측치 및 인접지역 관측치 정보 반영
  - kNN 알고리듬으로 최단거리 지역 k개 파악, k 지역 과거 미세먼지 오염도 정보 추정 반영

다층순환신경망



kNN 공간 순환신경망



데이터 전처리: 결측값 + GIS 보간

- 결측값 대치
  - 칼만 필터(Kalman filter)
    - 시계열 데이터 결측값 대치 방법론
    - 과거의 측정값과 새로운 측정값을 사용하여 모수의 평균을 추정
    - 실측값들 사이에 있는 결측값들을 선형으로 보간
- GIS 보간
  - 종관기상관측장비와 대기오염 측정소의 위치가 다르기 때문에 보간 실시
  - 역거리가중법(IDW)
    - 종관기상관측장비의 위경도 값을 대기오염 측정소의 거리에 반비례하게 가중치를 두어 보간점의 값을 계산

LSTM 구조 및 hyper parameter

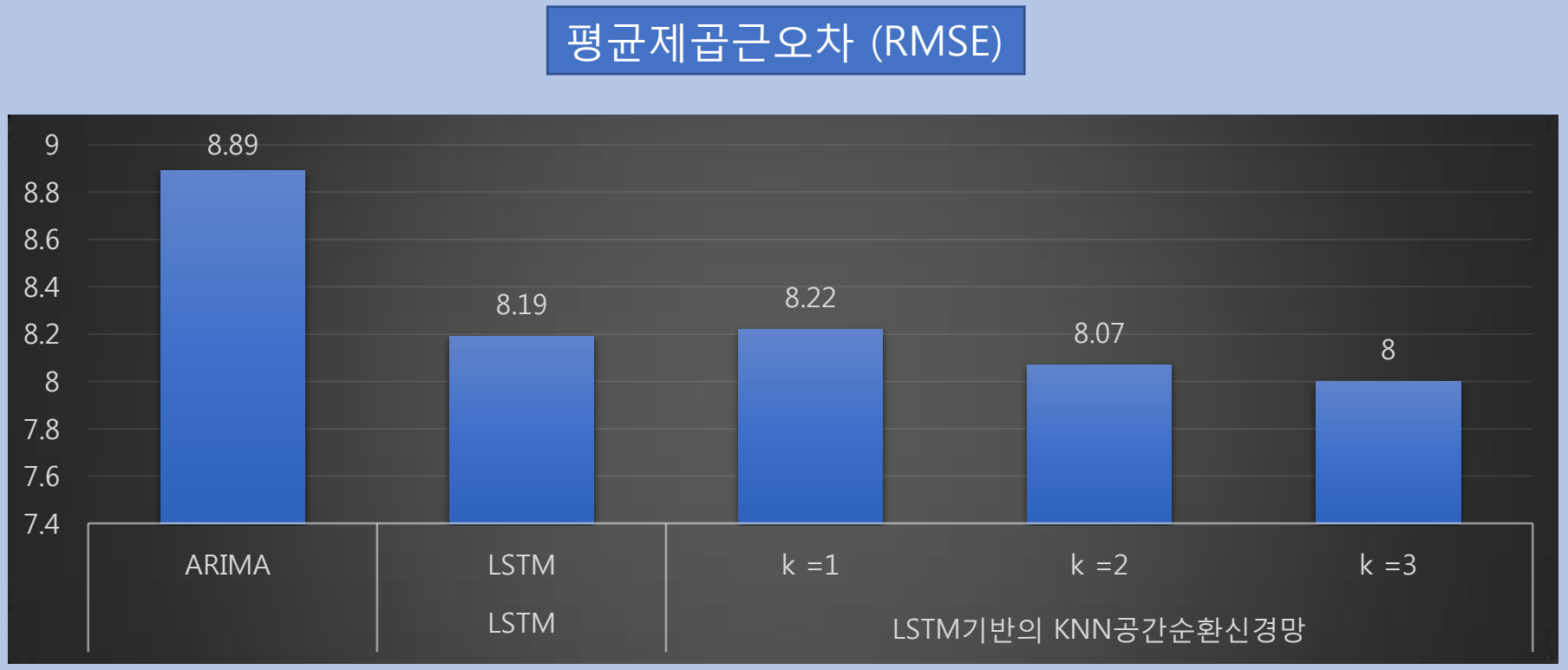
- 활성화 함수 : tanh(hyperbolic tangent)
- 손실 함수 : 평균 제곱 오차(MSE)
- 모형 학습 알고리즘 : RMSProp – 가중치 업데이트

Hyper Parameter

Time Step	뉴런 수	Batch 크기	학습률	Epochs
2	36	128	0.0005	120

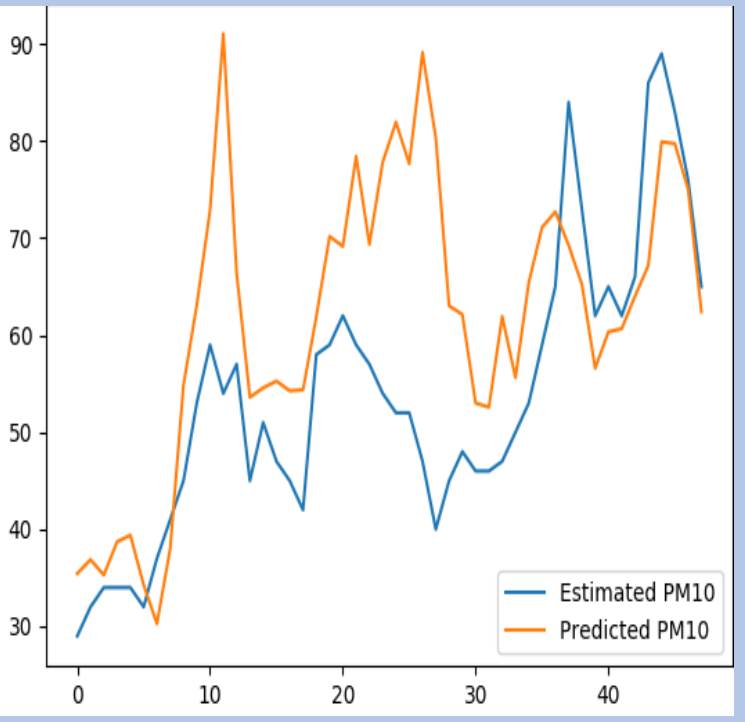
성과: ARIMA 대비 평균제곱근오차 10% 개선

- 평균제곱근 오차: ARIMA > LSTM > kNN 공간순환신경망
- 인근지역 3개 정보 반영한 kNN 공간순환신경망: 예측오차 최소화

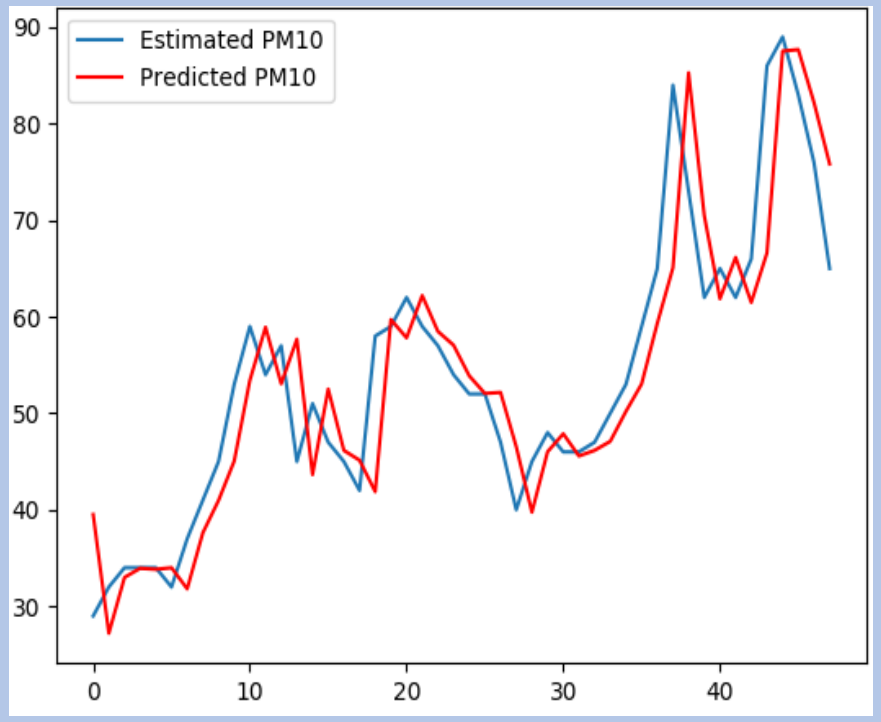


예측 정확도 비교 : 강남구 PM10 농도

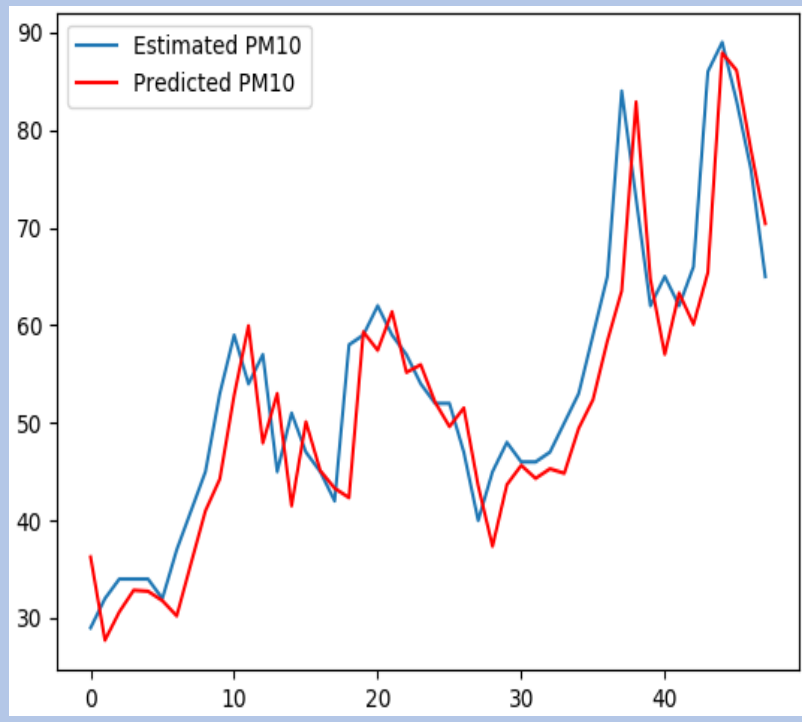
회귀분석



ARIMA



LSTM

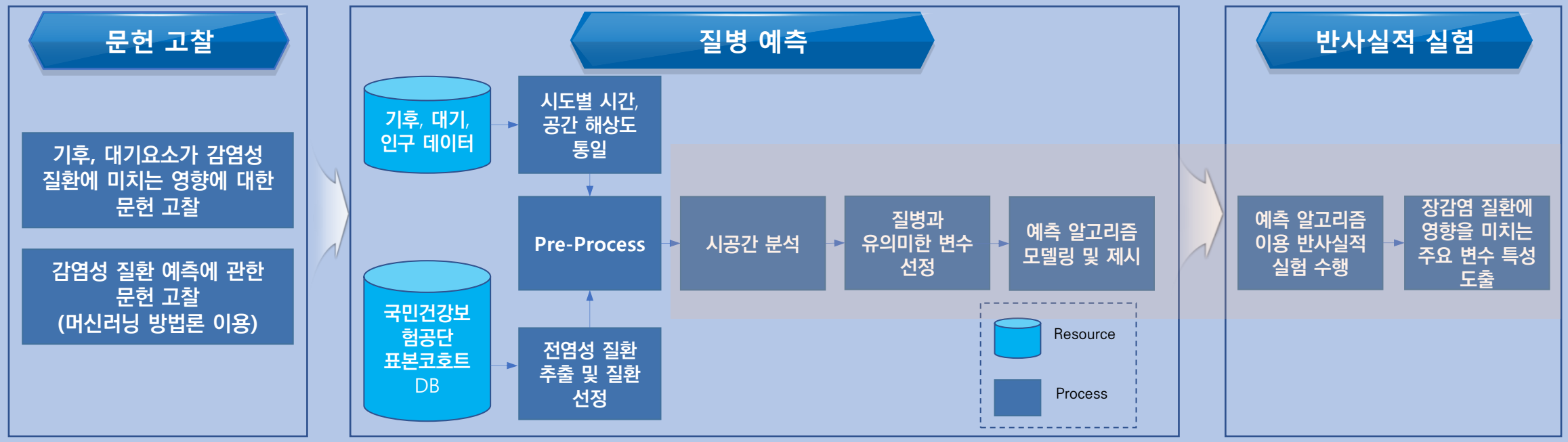


3. 연구성과 : 2017년 연구결과

2. 기후변화에 따른 전염성 질환 예측 (강선아)

기후변화에 따른 감염성 질환 예측

- 연구내용
  - 2009년~2013년(5개년)동안 발생하는 장감염 질환의 시공간 분석 및 예측 알고리즘 구축,
  - 반사실적 실험을 통해 장감염 질환에 영향을 미치는 주요 변수 파악
- 연구대상
  - 2009년부터 2013년까지 연속적으로 발생한 장감염 질환(국민건강보험공단 표본 코호트 DB 기준)



<그림 2-1> 연구 프로세스

데이터 전처리: 질병 선정 및 건수 도출

설명변수 데이터 전처리

- Step 1. 공간 해상도
  - 측정소 데이터는 공간적으로 점(point)데이터이고, 시군구/시도의 경우 면(polygon)데이터임
  - 공간해상도를 맞추기 위하여 같은 시군구/시도에 위치한 측정소의 데이터를 평균 내어 매칭
- Step 2. 시간 해상도
  - 시간해상도는 년, 월, 일 모두 다르며, 분석을 수행할 시간해상도는 월 단위
  - 월 단위보다 시간해상도가 낮은 경우: 시간, 일 단위 데이터의 평균을 월 단위 데이터로 사용
  - 월 단위보다 시간해상도가 높은 경우: 연 데이터를 12로 나누어 사용하고, 농도, 밀도(인구)와 같은 경우 연 데이터를 그대로 사용

질병 건수 산정 및 질병 선정

- Step 1. 자격 DB와 진료 DB 연계: 무진료 기간을 0으로 처리하여 질병 건수 산정
- Step 2. 진료 DB 주상병명과 부상병명이 다른 경우: 다른 케이스로 간주
- Step 3. 2009~2013년 연속 발생 질병만 분석 대상으로 고려
- Step 4. 질병코드(한국질병표준사인분류 기준) 소수점 그룹화 : 질병 레벨을 높여 건수를 산출

질환의 발생건수 및 월별 시계열 분석을 통해 장감염 질환을 분석대상 질병으로 선정

장감염 예측 알고리듬 구축

- 기상인자, 대기인자, 인구통계적 데이터 및 위, 경도 데이터 이용하여 장감염 질환 예측 알고리즘 구축
  - OLS 회귀분석, LASSO 회귀분석, 심층신경망 모형 구축
    - 알고리즘 학습을 위해 데이터를 학습 데이터와 테스트 데이터로 구분하였으며, 이 때 층화추출법을 사용함
      - 연도별, 월별, 시군구별로 데이터를 구분한 후 학습 데이터 70%, 테스트 데이터 30% 추출
    - 데이터의 범위를 통일시키기 위해 min-max 표준화를 수행

추정대상 및 설명변수

구분	변수 명
설명변수 (41개)	월(month), 위도, 경도, SO2_mean, CO_mean, O3_mean, NO2_mean, PM10_mean, SO2_max, CO_max, O3_max, NO2_max, PM10_max, 평균기온, 평균최고기온, 평균최저기온, 최고기온, 최저기온, 평균현지기압, 평균해면기압, 최고해면기압, 최저해면기압, 평균수증기압, 최고수증기압, 최저수증기압, 평균이슬점온도, 평균상대습도, 최소상대습도, 월합강수량, 일최다강수량, 평균풍속, 최대풍속, 최대순간풍속, 일조시간합, 일조율, 월적설량합, 평균.최저초상온도, 최저초상온도, 평균지면온도, 총인구수, 인구밀도
추정대상	시군구별 월별 장감염 질병 발생 건수

심층신경망 Hyper Parameter

추정대상	Parameter			
	Hidden Layer	Hidden node	Epoch	Activation
장감염 전체	3	500	30	ReLU
기타 세균성 장감염	3	500	7	ReLU
바이러스성 및 기타 감염성 장감염	3	500	70	ReLU

결과: 심층신경망 평균제곱근오차 10~25% 개선

- 장감염 질환 전체 예측: OLS/ LASSO 회귀분석보다 심층신경망의 성능이 대략 25% 향상
- 기타 세균성 장감염 및 바이러스성 및 기타 감염성 장감염 예측: OLS/ LASSO 회귀분석보다 심층신경망의 성능이 대략 10% 향상

장감염 질환 예측 표준제곱근오차 비교

추정 대상	OLS	LASSO	심층신경망
장감염 질환 전체	20.040	20.033	15.656
기타 세균성 장감염	4.342	4.342	4.071
바이러스성 및 기타 감염성 장감염	5.440	5.431	5.042

반사실적 실험: 주요 변수 양적 영향 파악

- 장감염 질환 발생에 영향을 미치는 주요 변수를 파악하기 위해 반사실적 실험 수행

- 반사실적 실험은 다음과 같은 방법으로 수행됨

테스트 데이터 셋의 설명변수가 i개 있다면, i개 설명변수에 대해 각각 10%씩 값을 상승시킴

$$x_i = (1 + 0.1)x_i$$

구축된 모델링(심층 신경망)에 X_i를 적용하여 장감염 발생빈도 변화율을 계산

$$yhat{\cdot}0 = DNN(x)$$

$$yhat{\cdot}1 = DNN(1.1x)$$

$$dy/y = \frac{1}{N}\sum\frac{yhat{\cdot}1 - yhat{\cdot}0}{yhat{\cdot}0}$$

반사실적 실험 결과: 기후변수 양적 영향 파악

- 지역, 인구의 영향이 크게 나타남
  - 위도, 경도 값을 변화시켰을 때 장감염 예측 건수가 가장 크게 변동하는 것으로 나타남
  - 총 인구 수가 증가하는 비율만큼 장감염 발생 증가
- 전체 장감염: [촉진요인] 현지기압, 일조율, 평균최저초상온도 [억제요인] 평균지면온도, 최저기온
- 세균성 장감염: [촉진요인] 평균현지기압, 평균최저초상온도, 최저초상온도, 오존, 일산화탄소
- 바이러스성 및 기타 감염성 장감염:[촉진요인] 일조율, 일조시간 [억제요인] 평균현지기압, 최고/최저/평균 해면기압, 평균/최소 상대습도

반사실적 실험 결과

변인	장감염 전체	변인	기타세균성장감염	변인	바이러스 및 기타 감염성 장감염
평균현지기압	5.1	평균현지기압	2.78	일조율	13.59
일조율	3.05	평균최저초상온도	1.67	일조시간합	8.99
평균최저초상온도	1.54	최저초상온도	1.63	최소상대습도	-2.58
평균최고기온	1.51	O3_max	1.45	평균상대습도	-4.32
평균기온	0.18	CO_mean	1.17	평균해면기압	-8.36
월적설량합	0	평균기온	-0.37	최고해면기압	-8.76
최저초상온도	-0.04	평균최고기온	-0.57	최저해면기압	-35.19
평균최저기온	-0.63	평균최저기온	-0.85	평균현지기압	-61.34
최고기온	-0.92				
최저기온	-1.2				
평균지면온도	-2.37				

결론

장감염 예측 결과

- 전체 장감염: OLS/ LASSO 회귀분석 RMSE 20.040, 20.033 → 심층신경망 RMSE 15.656, 예측오차 25% 감소
- 기타 세균성 장감염: OLS/ LASSO 회귀분석 RMSE 4.342 → 심층신경망 RMSE 4.071, 예측오차 10% 감소
- 바이러스 및 기타감염성 장감염: OLS/LASSO의회귀분석 RMSE 5.440, 5.431 → 심층신경망 RMSE 5.042, 예측오차 10% 감소

반사실적 실험 결과

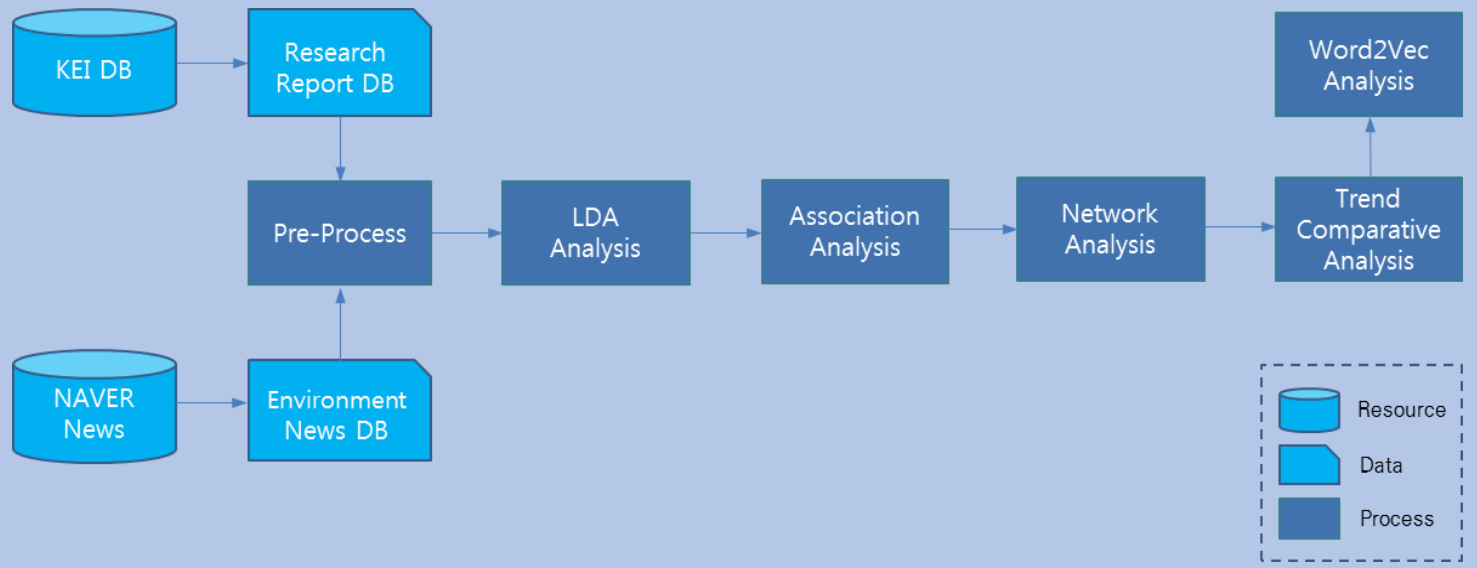
- 장감염 질환 전체: 평균최고기온, 최저해면기압, 평균수증기압, 일조율, 평균최저초상온도 등 기압/기온 관련 변수 영향이 큼
- 기타 세균성 장감염: 평균기압, 초상온도, 대기오염물질(오존,일산화탄소) 오염도의 영향이 큼
- 바이러스 및 기타감염성 장감염: 일조율/상대습도/기압의 영향이 큼

3. 연구결과: 2017년 연구성과

3. 텍스트마이닝을 이용한 KEI 연구동향분석 (김도연)

텍스트마이닝을 이용한 KEI 연구동향분석

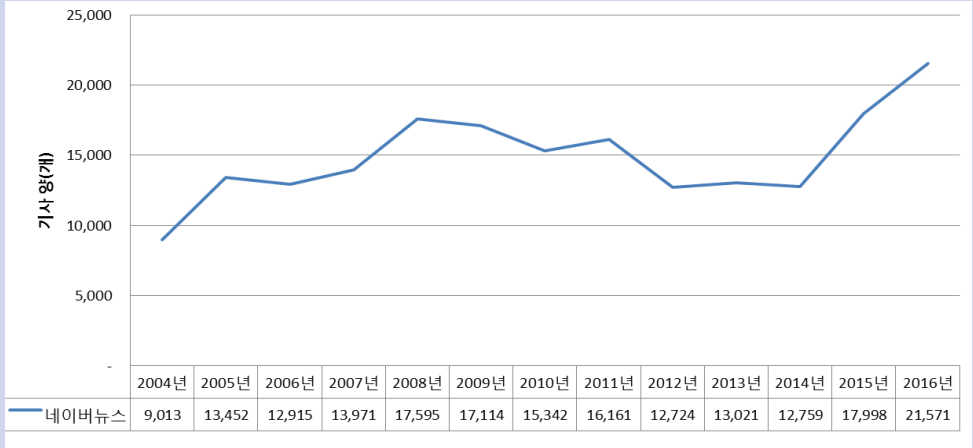
- 연구 내용
  - KEI 연구보고서(1993-2016)와 NAVER 환경뉴스(2004-2016) 데이터 이용
    LDA(Latent Dirichlet Allocation) 분석, 연관어 분석, 언어 네트워크 분석, word2vec 분석 수행
- 연구 프로세스 :



  - LDA Analysis: 매체 별 주요 토픽을 추출하고 이슈 변화를 관찰
  - Association & Network Analysis: 매체 별 중심 키워드 파악 및 환경 분야 이슈 발굴
  - Word2vec Analysis 분석: 환경 분야 이슈 관련 키워드 간의 문장 내 관계 분석

데이터 수집 및 전처리

구분	내용
수집 도구	Java HTML Parser - jsoup
산출 조건	네이버 뉴스 -> 사회 분야 -> 환경 분야
산출 기간	2004-01-01 00:00:00 ~ 2016-12-12 23:59:59 (총 13개년)
산출 영역	제목, 날짜(년, 월, 일, 시간), 언론사
산출 유형	지면기사, 보도자료
언론사	총 101개 (부록2 참조)
산출 양	**총 193,636개**



	2004년	2005년	2006년	2007년	2008년	2009년	2010년	2011년	2012년	2013년	2014년	2015년	2016년
네이버뉴스	9,013	13,452	12,915	13,971	17,595	17,114	15,342	16,161	12,724	13,021	12,759	17,998	21,571

텍스트 데이터 전처리 과정

1. 형태소분석기 수행
   - R에서 제공하는 형태소분석기 패키지 (KoNLP)를 사용하여 명사 추출
2. 특수문자, 특정단어 등 불용어 삭제
3. 단어길이 한글자 삭제
4. 출현빈도가 매우 낮은 단어(Sparse Terms) 삭제
5. Low TF-IDF 단어 삭제
6. DTM(Document Term Matrix)형태로 변형

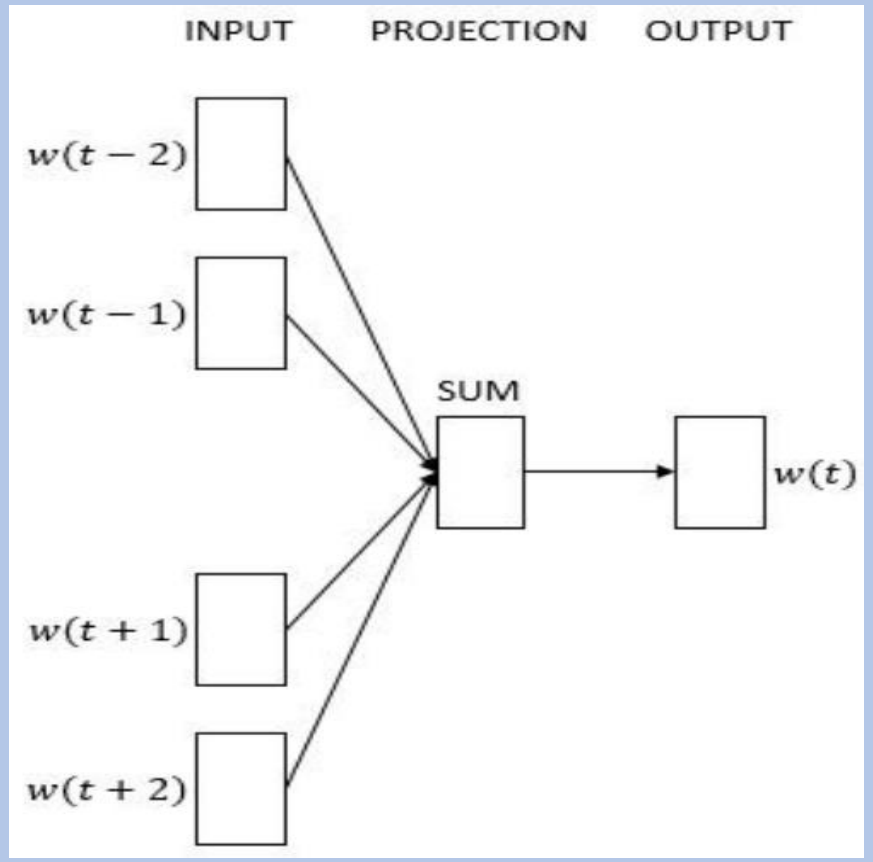
예)

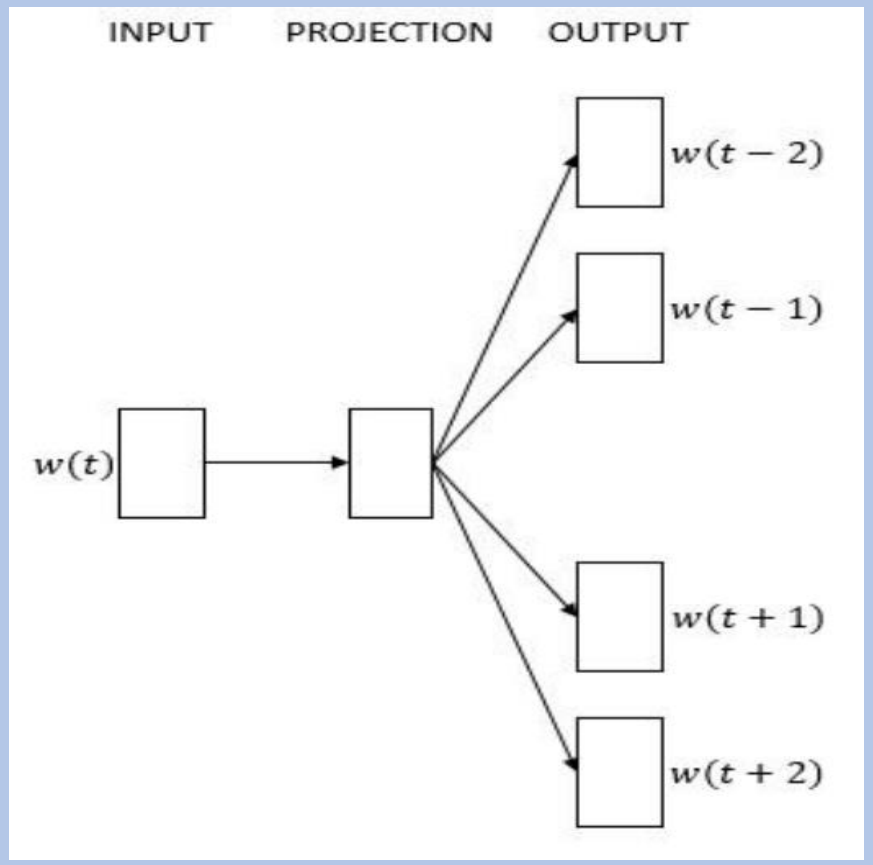
Term / Document	기후	오염	…	한강
보고서 1	0	2	…	1
보고서 2	2	1	…	1
…				…
보고서3	0	1	…	5

텍스트 마이닝 방법론

- LDA : 분석 대상 문서 단어 분포 기반 문서 토픽 추출
- 키워드 연관성 분석 : 동시 출현 빈도가 높은 단어 조합 파악
- 키워드 네트워크 분석: 동시 출현 빈도 이용 단어 간 연결 관계 파악
  - 밀도: 네트워크 내 연결점 간의 관계가 나타나는 빈도
  - 중심성: 다른 연결점과의 연결 정도
- Word2vec: 문장 내 단어 간 관계 파악
  - CBOW(Continuous Bag of Words) : 주변 단어를 이용하여 특정 단어를 예측
  - Skip-gram: 특정 단어를 이용하여 주변 단어를 예측

Word2Vec : CBOW , Skip-gram

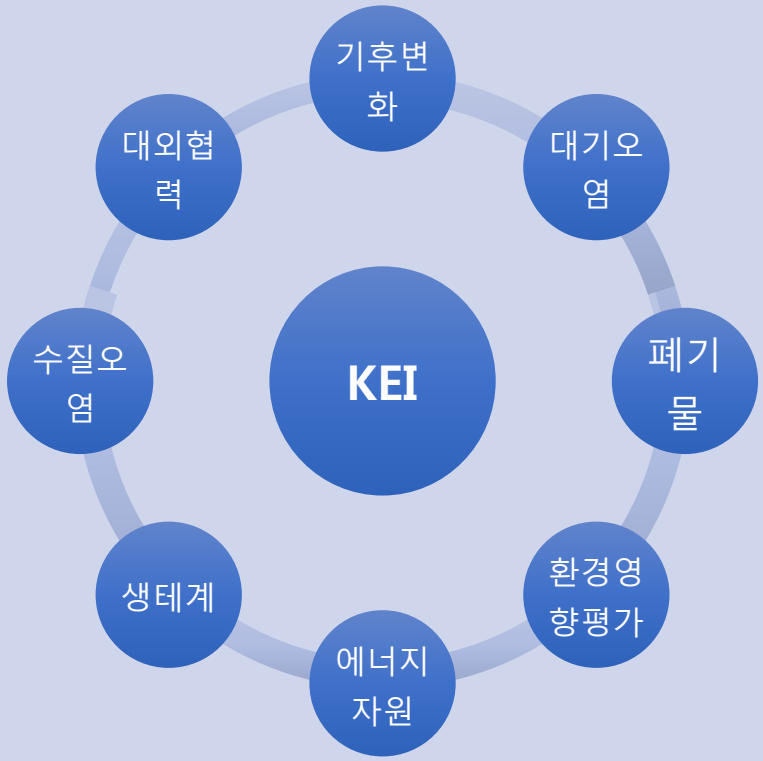




LDA 분석: 매체 별 토픽 구성

- 매체별 13개년(2004-2016) 전체 데이터 분석
  - 공통 토픽 : 기후변화, 폐기물, 환경영향평가, 에너지자원, 수질오염, 대외협력
  - 기타 토픽 : KEI 연구보고서(생태계), NAVER 환경뉴스(유전자 변형/소음, 해양/풍력, 보건/데이터)

KEI 연구보고서 (2004~2016)	네이버 환경뉴스(2004~2016)

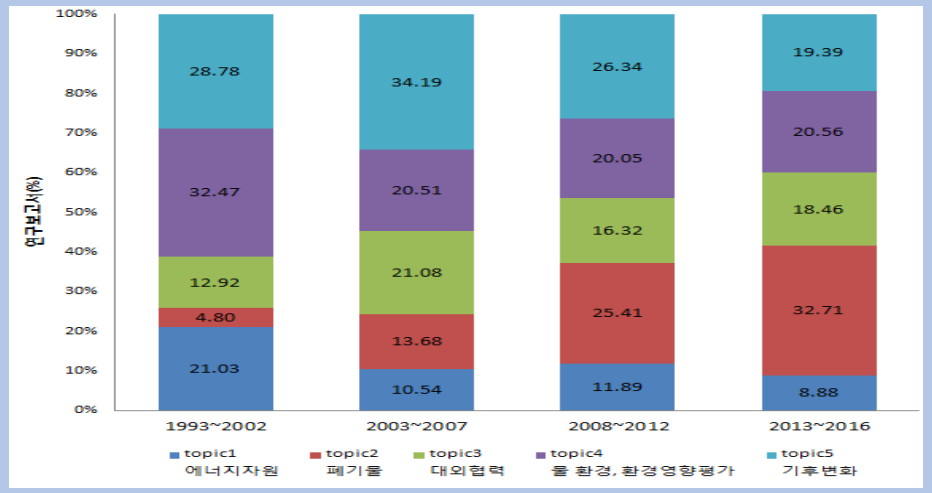


LDA 분석: 기간 별 토픽 구성 변화

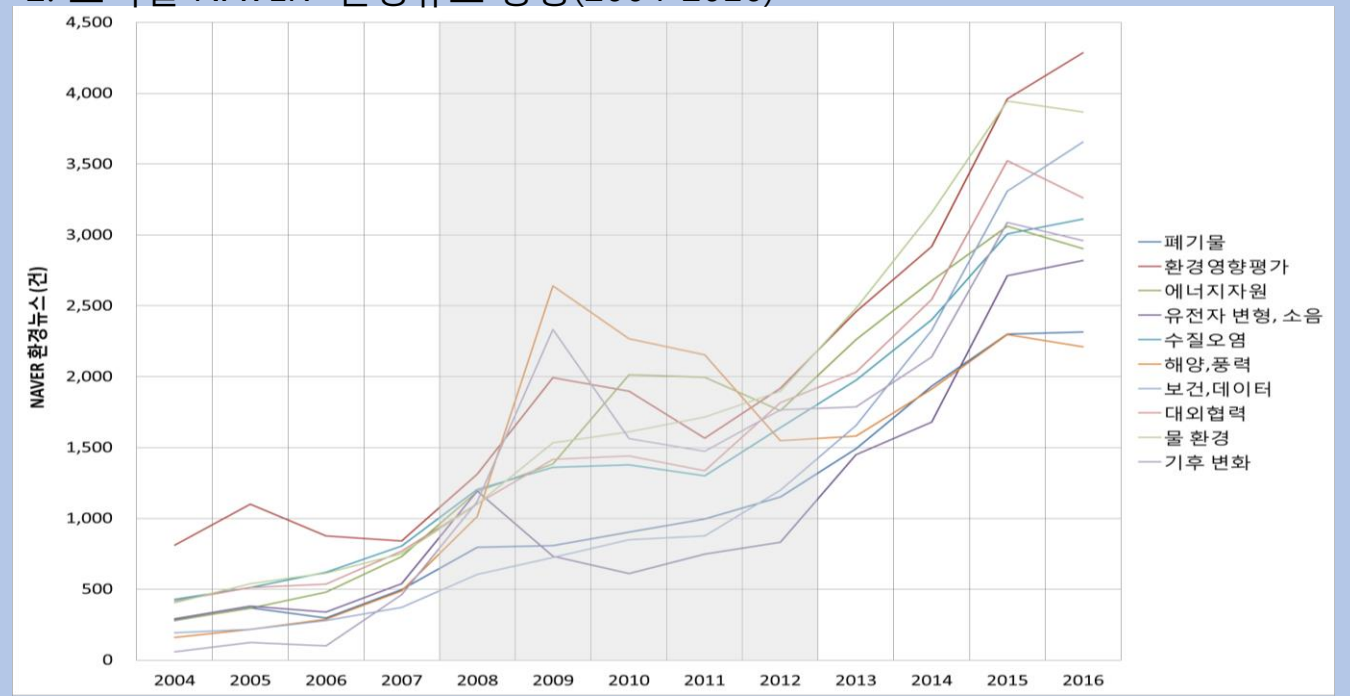
- KEI 연구보고서
  - 1구간(1993~ 2002년) : 전반적으로 토픽 별 연구추세 유사
  - 2구간(2003~ 2007년) : 기후변화 관련 연구 활발
  - 3구간(2008~ 2012년) : 폐기물, 물 환경/환경영향평가 연구가 급증
  - 4구간(2013~ 2016년) : 폐기물 연구 활발, 2015년 기점 연구의 양 감소
- NAVER 환경뉴스
  - 2구간(2008~2012년)을 제외하고 기사 양이 증가하는 추세를 보임
    - 2008년은 '유전자 변형/소음' 관련 기사가 급증함
    - 2009년은 '해양/풍력', '기후변화', '환경영향평가' 관련 기사가 급증함
- KEI 연구: '기후변화' 선도/ '수질오염' 후행/ '에너지-자원' 부족
  - 최근 '유전자 변형, 소음' '보건, 데이터' 환경 뉴스 증가 추세가 반영되지 못함

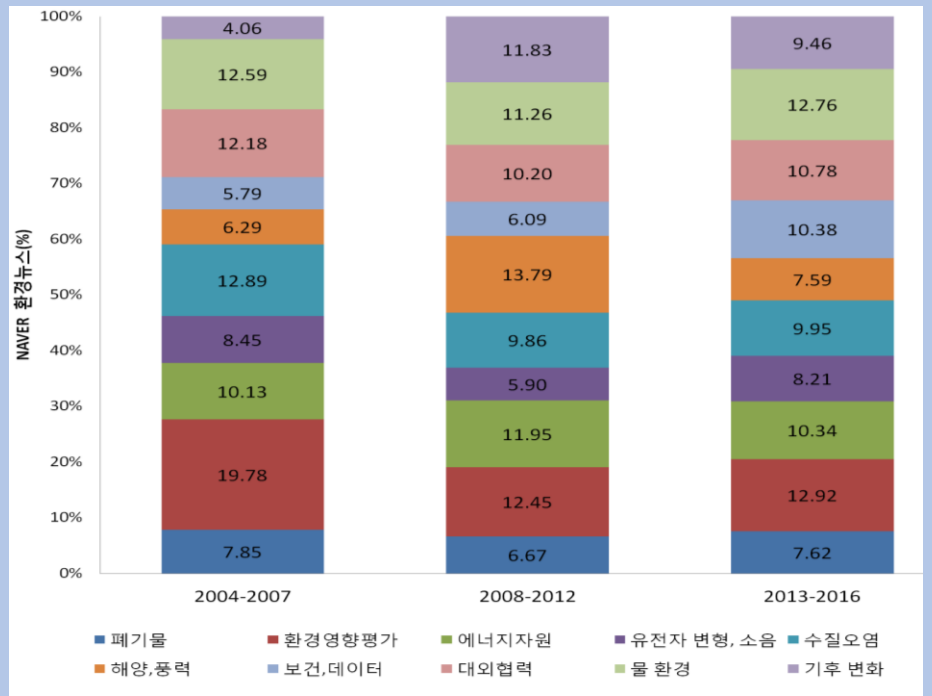
기간 별 토픽 동향 변화

1. 토픽별 KEI 연구보고서 동향(1993-2016)



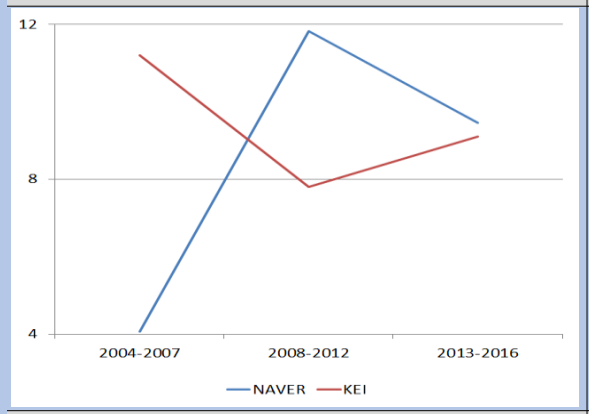
2. 토픽별 NAVER 환경뉴스 동향(2004-2016)



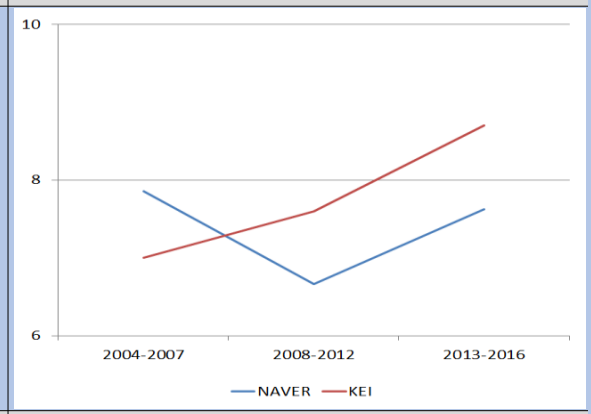


4. 공통 토픽 동향 비교

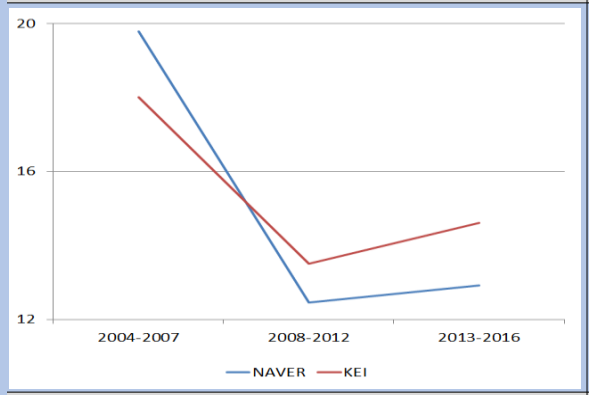
1. 기후변화(대기)



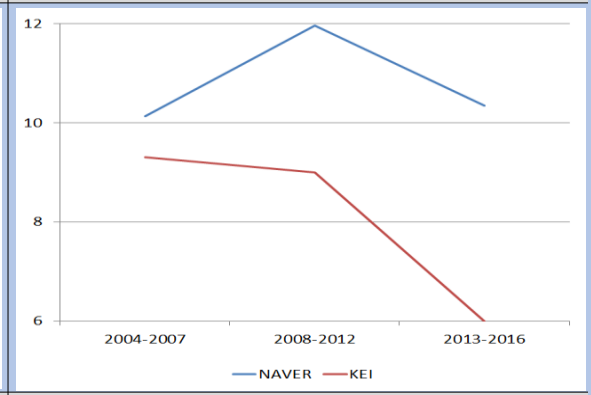
2. 폐기물



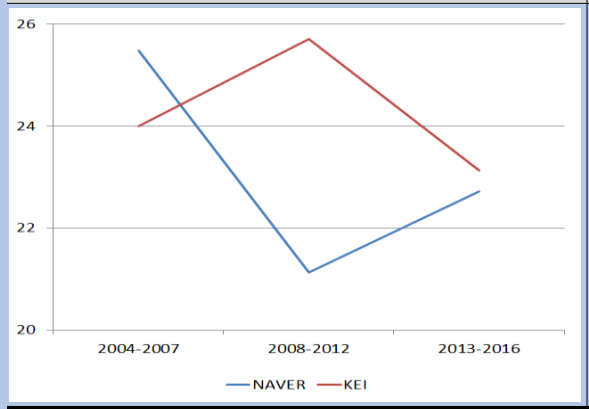
3. 환경영향평가



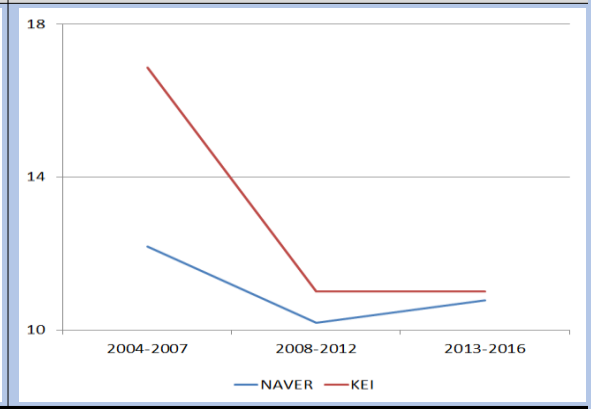
4. 에너지 자원



5. 수질오염(물환경)

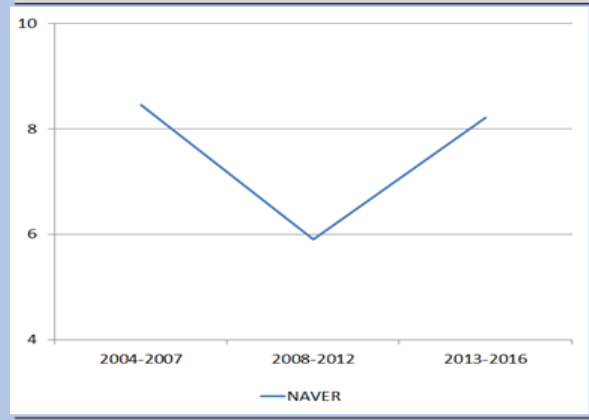


6. 대외협력

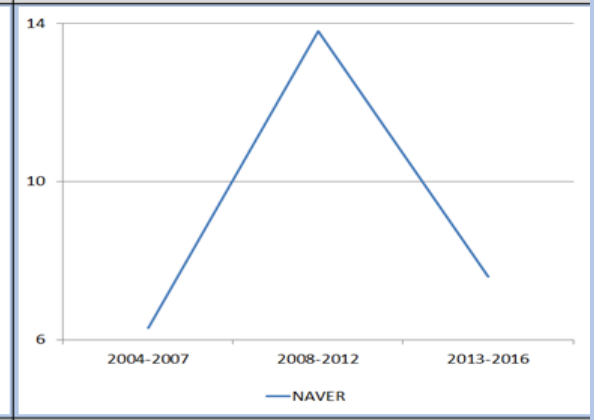


5. 기타 토픽 동향

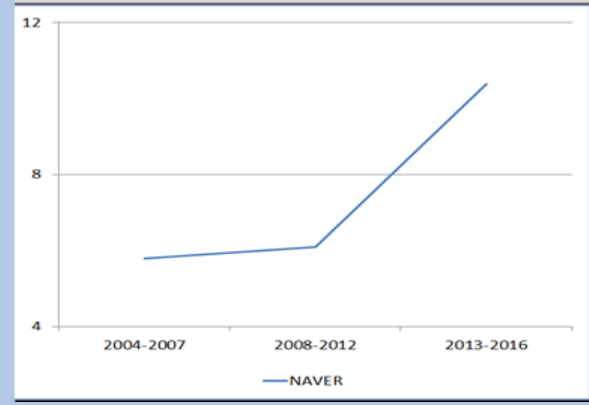
유전자 변형, 소음



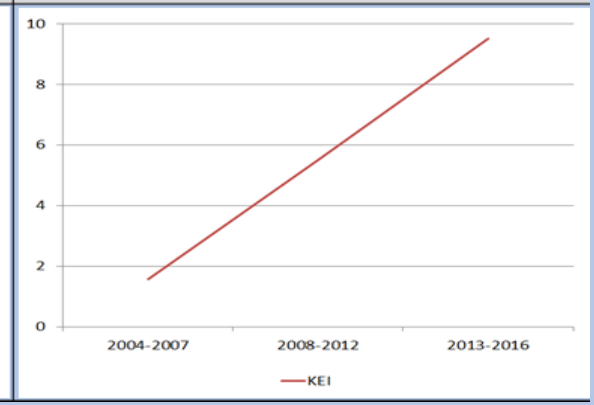
해양, 풍력



보건, 데이터



생태계



- **토픽 동향 유사:** '환경영향평가', '대외협력'
- **토픽 동향 차이**: KEI '기후변화' 선도, '수질오염' 후행, '에너지-자원' 부족
- **최근** '유전자 변형, 소음'과 '보건, 데이터' 관련 환경 이슈가 대두

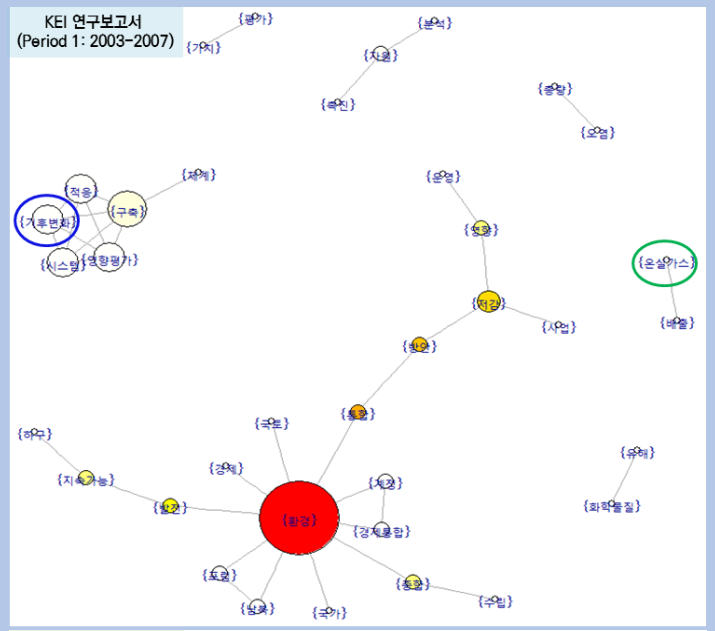
키워드 네트워크 분석: 기간별 네트워크 변화

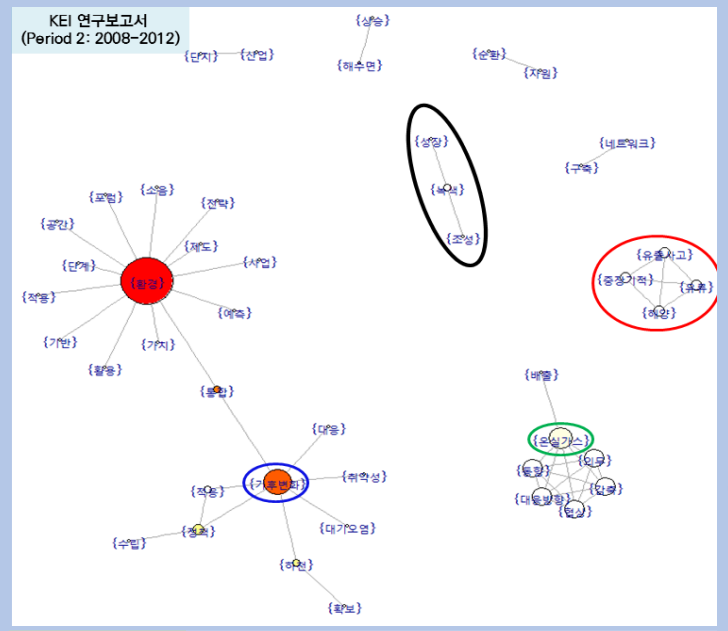
- 기후변화 : KEI 가 Trend를 선도하였으나 최근 추세 반영 미흡
  - 2003~ 07년 KEI 연구보고서 기후변화 네트워크 생성 → 2008~12년 NAVER 환경뉴스 기후변화 네트워크 선도
  - 2008~16 년 KEI 연구보고서 기후변화 매개중심성 강화 vs. NAVER 환경뉴스 기후변화 관련 키워드 네트워크 분화 (태풍, 한파, 대설)
- 사건사고: 해양 오염 관련 Trend 는 NAVER 선도
  - 2003~07년 NAVER 뉴스 해양오염 키워드 네트워크 생성 : 2008~12 KEI 연구보고서 해양오염 키워드 네트워크 생성

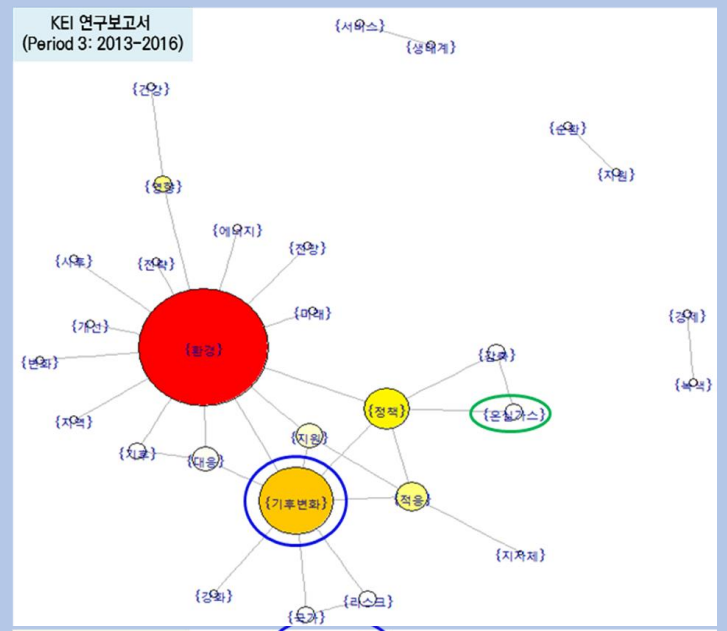
키워드 네트워크 변화

1. 기후변화 2. 온실가스 3. 태안 기름 유출 사고 4. 녹색성장

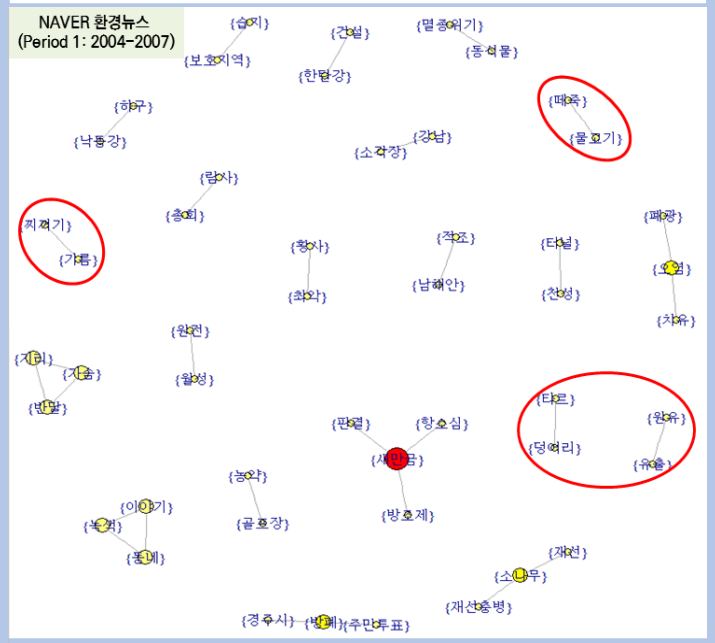
KEI 연구보고서

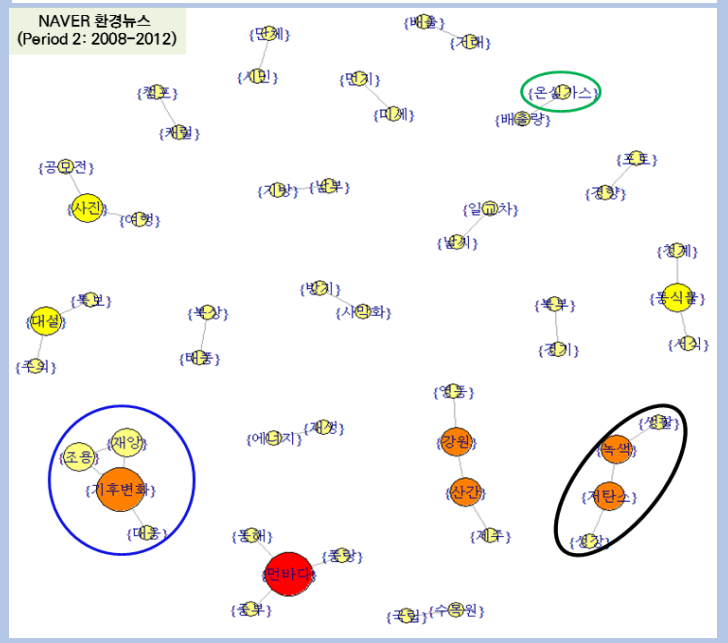


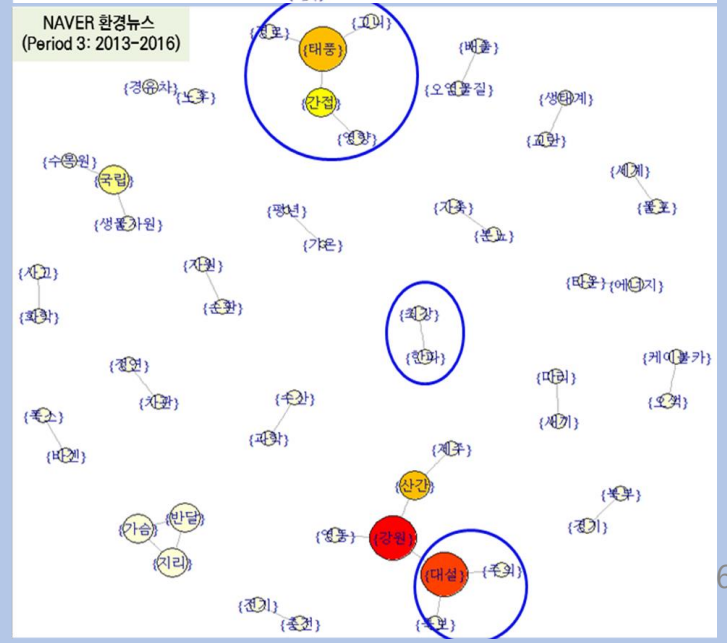




NAVER 환경 뉴스





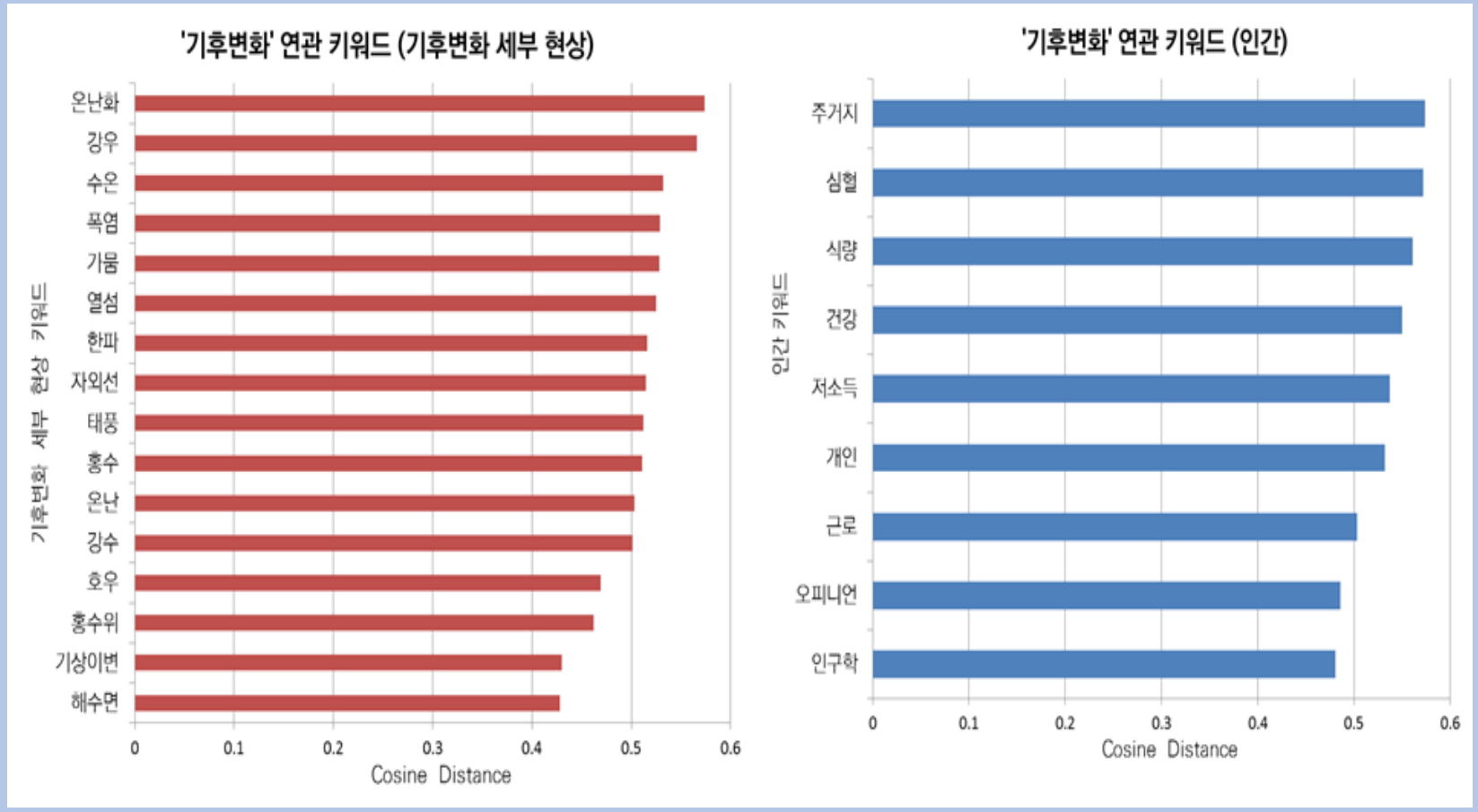


Word2Vec 분석: Keyword 간 문장 내 관계

- Skip-gram 모델: ‘기후변화’ 관련 키워드를 넣어 주변에 있는 단어들을 예측
- 기후변화 세부 현상 키워드: ’기후변화’ Skip gram 분석 결과→ ‘온난화’, ‘홍수’, ‘가뭄’ 선정
  - 3가지 키워드에 대한 Word2Vec 분석을 매체별 수행 후 비교 분석
- 분석 결과 : 연구보고서(국민의 삶의 질) 와 뉴스(관심사)의 강조점 차이 발견
  - 온난화: 인간 관련 단어(KEI) vs. 생물 및 식량 관련 단어(NAVER)
  - 홍수: 대한민국 지역 (KEI) vs. 중국 지역(NAVER)
  - 가뭄: 인간 관련 단어(KEI) vs. 농업 관련 단어(NAVER)

기후변화 Skip-gram: 기후/인간 관련 단어

- KEI 연구보고서 자료 활용
- 기후변화 세부현상 지칭 단어 및 인간 관련 단어가 기후변화와 관계
- 기후변화 세부현상 지칭 단어 중 '온난화','홍수','가뭄' 선정
  - '태풍' : 태풍 이름 관련 단어가 관계 깊은 단어로 자주 출현 → 제외
  - '한파','폭설','폭염','폭우': 기상 관련 불용어(33도, 영하, 곳곳, 오후) 가 관계 깊은 단어로 자주 출현 → 제외

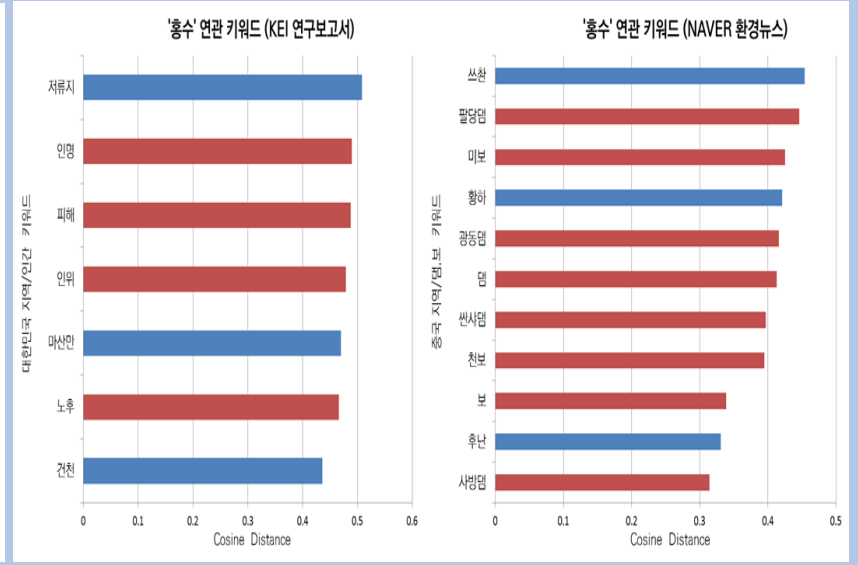


Word2Vec 분석: Keyword 간 문장 내 관계 분석

1. 온난화

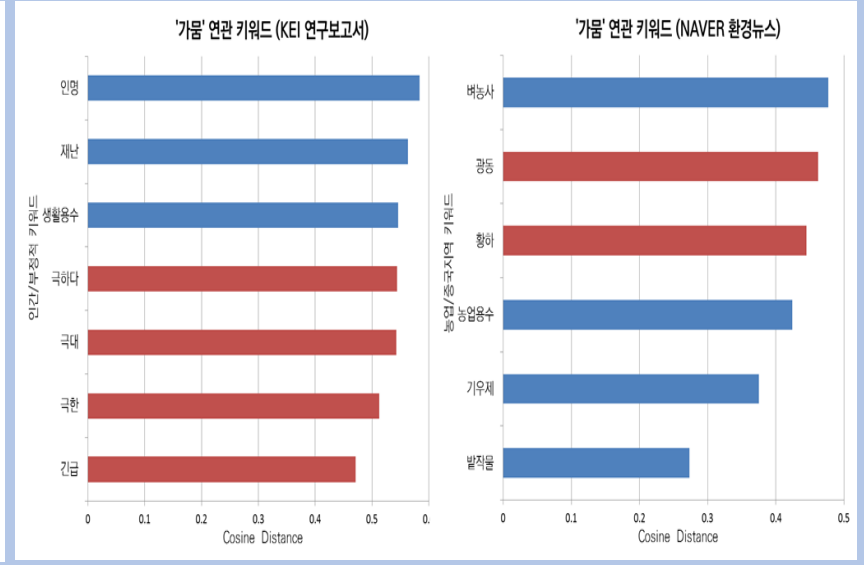
- KEI 연구보고서:
  인간 키워드(거주, 인류, 인명, 호흡기계, 심혈) 통해 온난화 관련 환경연구는 호흡기계 질환 및 심혈관계 질환과 관련된 연구가 활발히 진행되었음을 확인함
- NAVER 환경뉴스:
  생물 및 식량 키워드(플랑크톤, 산호초, 말매미, 식량) 통해 온난화의 영향을 많이 받는 플랑크톤, 산호초, 말매미와 관련된 환경문제 이슈가 대두되었음을 확인함

2. 홍수



- KEI 연구보고서:
  대한민국 지역 키워드(건천, 마산만) 통해 홍수 피해를 입은 국내 지역에 관한 환경연구가 활발히 진행되었음을 확인함
- NAVER 환경뉴스:
  중국 지역 키워드(후난, 황하, 쓰촨) 통해 환경뉴스의 관심사는 대규모 홍수 피해의 심각함에 중점을 두고 있음을 파악함

3. 가뭄



- KEI 연구보고서:
  인간 키워드(생활용수, 재난, 인명) 통해 환경 연구의 관심사는 가뭄으로 인한 생활용수 부족 문제에 중점을 두고 있음을 파악함
- NAVER 환경뉴스:
  농업 키워드(밭작물, 기우제, 농업용수, 벼농사) 통해 환경뉴스의 관심사는 가뭄으로 인한 농업용수 부족 문제에 중점을 두고 있음을 파악함

결론

LDA 분석 결과

- NAVER 환경뉴스와 KEI 연구보고서 전체 데이터(2004-2016)를 비교 분석한 결과 공통적으로 '기후변화', '폐기물', '환경영향평가', '에너지자원', '수질오염', '대외협력' 토픽을 중요하게 다루고 있음을 확인
- **환경뉴스는 '보건/데이터', '유전자 변형/소음' 관련 환경기사가 최근 많이 보도 되어 향후 성장 가능성 있는 연구 주제로 판단**

연관어 및 네트워크 분석 결과

- 환경뉴스는 기후변화의 세분화된 주제를 중심으로 보도가 되고 있으며, 환경연구는 기후변화 일반을 중심으로 연구되었음을 확인
- **따라서 향후 기후변화의 세부주제에 대한 연구가 유망한 연구 주제로 파악됨**
- 포괄적인 관점에서 본 연구는 환경연구의 주제가 환경뉴스의 주제를 시차를 두고 추종하는 경향이 있음을 확인함
- **따라서 이러한 시차를 메울 수 있는 연구에 대한 요구가 앞으로도 지속될 전망**

Word2Vec 분석 결과

- 환경연구의 관심사는 국민의 삶의 질에 중점을 두고 있는 반면, 환경뉴스의 관심사는 기후변화로 인한 피해의 심각함에 중점
- **정책 성과 제고를 목적으로 하는 환경연구문헌과 사실보도를 목적으로 하는 언론 간의 매체의 차이를 반영**

3. 연구결과: 2017년 연구성과

4. 미세먼지 오염도-발생요인 패턴분석 (김진형)

미세먼지 오염도-발생요인 패턴분석

- 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
  - 반사실적 분석을 통한 변수 영향 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

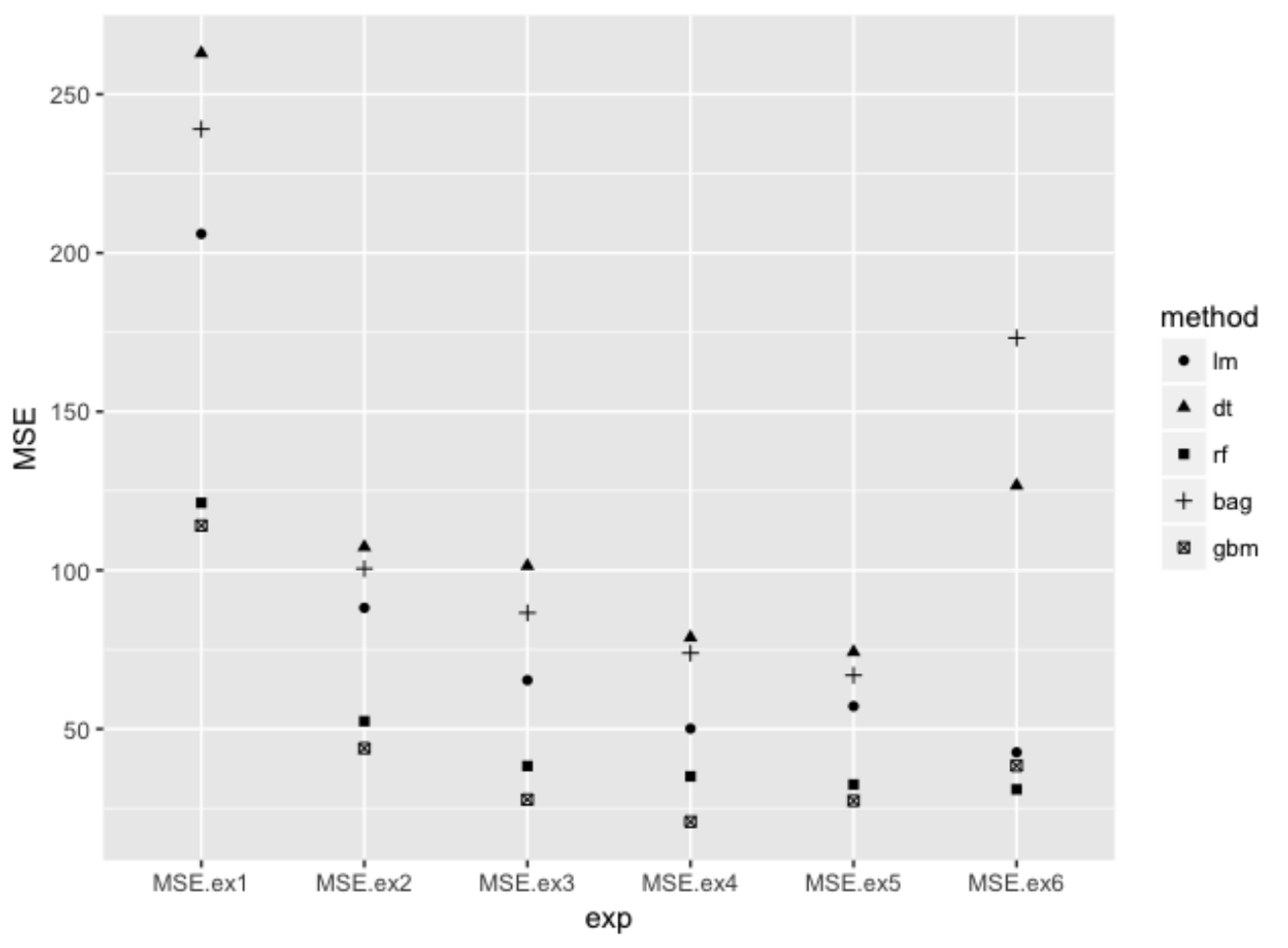
해당 시군구에 매칭

의사결정나무 추정 실험설계

- 데이터 별 구축기간 차이로 인해 실험 기간을 달리하여 모델 구축
- 대기오염물질 배출량: 2007년 대분류 변경 이전과 이후로 나누어 실험
  - 실험1: 대분류 변경 이전/ 실험2~6의 경우 대분류 변경 이후
- 중국 대기질 데이터의 경우 구축 시기에 따라 기간을 나누어 실험
  - 실험 1,2: 중국 대기질 데이터를 포함하지 않음
  - 실험 3~6: 베이징 PM2.5 데이터를 포함함
  - 실험 4~6: 상하이 PM2.5 데이터를 포함함
  - 실험 5: 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

분석 결과: 예측오차



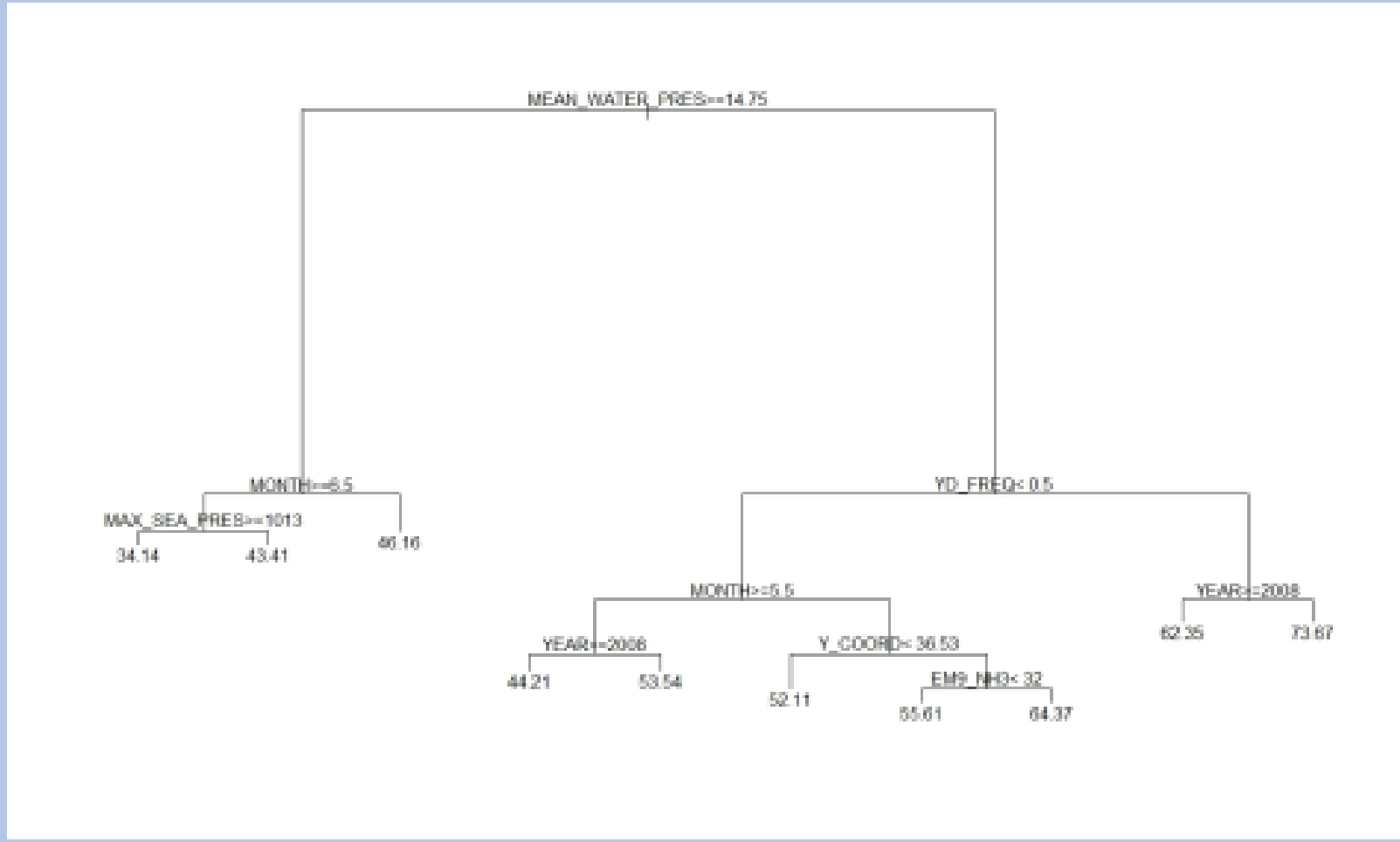
- 방법론 별 모델 정확도 비교
  - boosting 모델의 정확도가 가장 높고, random forest 모델이 선형회귀보다 정확도가 높음
  - Boosting 모델: 선형회귀 평균제곱오차 45.5% 감소 (6개 실험 평균)
  - random forest : 선형회귀 평균제곱오차 37.2% 감소 (6개 실험 평균)
    - 의사결정나무 /bagging 모델: 선형회귀보다 정확도가 악화, 예측 사용은 부적합
- 실험별 모델 정확도 비교
  - 실험 3~6 (의사결정나무와 bagging 결과 제외)의 정확도가 높음:
    - 중국 대기질 데이터를 이용하는 것이 모델 정확도를 높이는 것으로 판단됨

의사결정나무 변수 중요도: 분기점

- 분기점에 나타난 변수 : 중국 대기질의 영향을 반영
  - 황사관측일수, 베이징PM2.5표준값, 상하이 PM2.5

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

의사결정나무 분기점


MEAN_WATER_PRES>=14.75
MONTH>=6.5
MAX_SEA_PRES>=1013
34.14
43.41
46.16
YD_FREQ< 0.5
MONTH>=5.5
YEAR>=2008
62.35
73.87
YEAR>=2008
Y_COORD>= 36.53
44.21
53.54
52.11
EM9_NH3< 32
55.61
64.37

의사결정나무 변수중요도 (24개 모형)

- 6개 실험 4개 모형(의사결정 나무, Bagging, Random Forest, Boosting)에서 평균제곱오차 감소에 기여한 변수를 파악

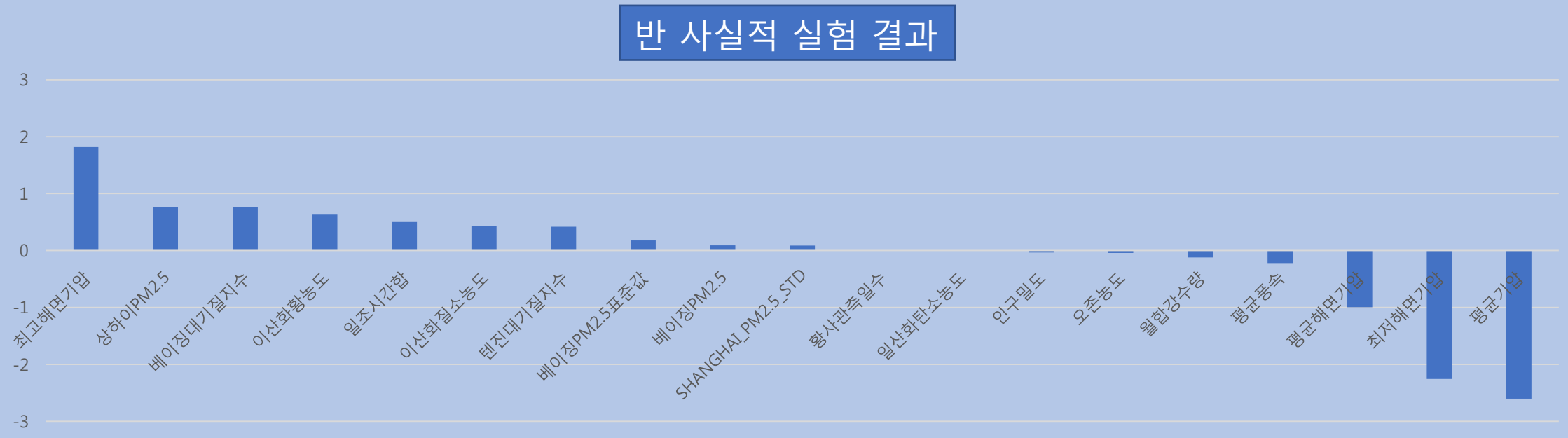
- 중국관련 변수: 황사, 대기오염
  - 실험 4 제외 5개 모델: 황사관측일
  - 실험 3~6 : 중국 대기질 관련 변수
    - 베이징PM2.5, 베이징PM2.5표준값, 상해PM2.5, 상해PM2.5표준값. 베이징대기질지수(AQI), 텐진대기질지수(AQI) 중요변수 선정

- 2차 미세먼지: NO2(이산화질소농도), SO2(이산화황농도), CO(일산화탄소농도), O3(오존농도)

- 기상기후요인 : 최고/평균/최저 해면기압, 월 강수량, 일조시간합, 평균풍속
  - 일조시간: 2차 미세먼지 생성 기여

반사실적 실험 결과: 기압/중국 대기질 영향

- 예측 정확도가 높은 4개의 모델: 실험 3~5 Boosting , 실험 6 Random forest
- 각 변수 10% 증가 시, 모델 예측 PM10 농도 변화 계산
  - 영향이 큰 변수(1) : 최고해면기압, 상하이PM2.5, 베이징대기질지수 , 텐진대기질 지수
  - 영향이 큰 변수(2): 이산화황 농도, 이산화질소 농도



결론 및 정책적 제언

1. 중국 대기질 영향
   - 중국의 대기질 데이터를 사용하면 모델의 정확도가 높아짐
     - 베이징, 상하이, 텐진 시의 대기질: 변수 중요도 상위
     - 황사관측일수: 대부분 모델 변수 중요도 측정에서 상위에 위치

2. 국내 NOx, SOx 관리 필요성: 2차 미세먼지 발생 억제 필요
   - NOx와 SOx: 높은 변수 중요도, 큰 정량적 효과
   - 일사시간합: 높은 변수 중요도, 2차 미세먼지 발생 촉진과 관련

3. 연구결과: 2017년 연구성과

5. 환경분야 빅데이터 수집방법연구 (한국진)

환경분야 빅데이터 수집방법연구

- 목적: 데이터 중심 연구 패러다임 대응을 위한 빅데이터 수집방법 연구
- 연구내용: 환경 데이터 수요 조사 → 수집 데이터 선정 → 수집 자동화
  - 수요 조사: 공공데이터 포털 활용신청 순위/ 원내데이터 활용사례조사/연구자 면담
  - 선정 대상: 공공데이터 포털 활용신청 순위 기준 기상청/한국환경공단 사례 선정
  - 수집 자동화 코드 : Web 게시 자료 → 수집 → RDBMS 형태로 전환 → 저장
- 수집 자동화 코드 구축: Python Beautifulsoup, json, pandas
  - [공공 데이터 포털] 기상청 동네예보정보조회서비스(최근 24시간) (통합 코드 4건+)
    - 여러 항목의 데이터를 한꺼번에 수집해서 후처리 과정에서 분리
  - 기상자료개방포털(코드 3건+)
    - 데이터 항목 별로 따로 따로 수집 가능, 후처리 과정이 복잡
  - [공공 데이터 포털] 한국환경공단 대기오염정보 조회 서비스 (통합코드 11건+)
    - 여러 항목의 데이터를 한꺼번에 수집해서 후처리 과정에서 분리
  - 에어코리아 (코드 20건+)
    - 데이터 항목 별로 따로 수집 가능, 후처리 간단

환경분야 빅데이터 수집사례(1/4)

- (공데 포털) 기상청 동네예보정보조회서비스 | 오픈API(JSON)

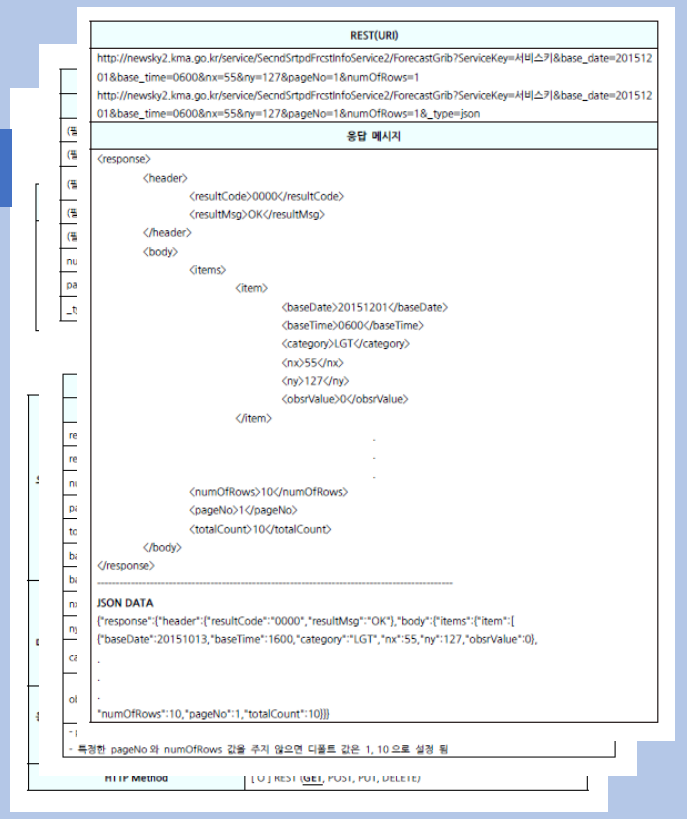
1. 초단기실황조회	2. 초단계예보조회	3. 동네예보조회	4. 예보버전조회

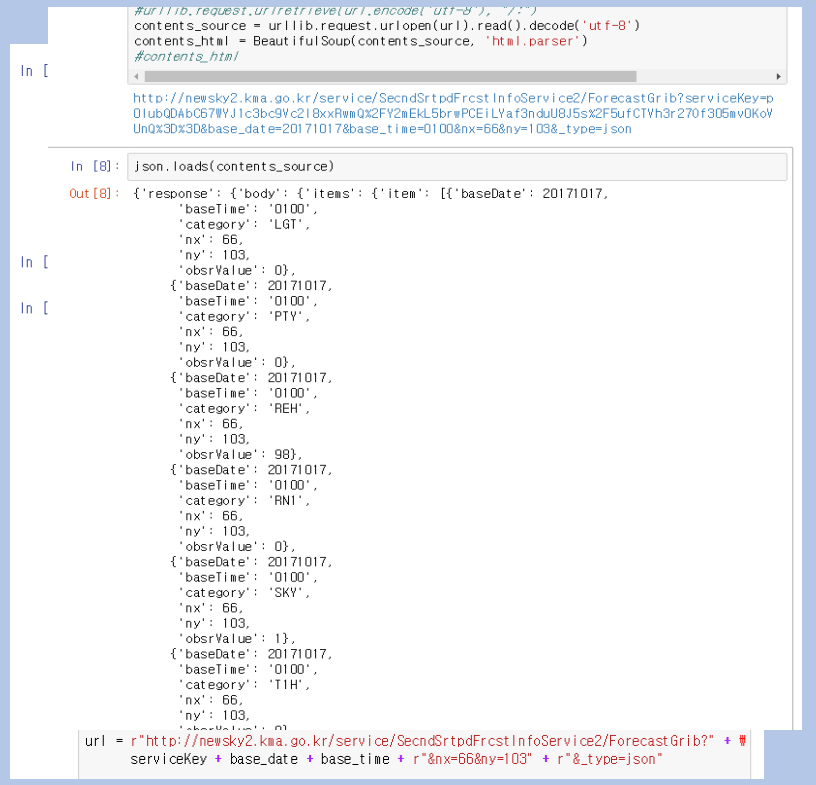
데이터 식별 / 데이터 서비스 탐색

메타데이터 확인 / 오픈API 분석

자동화 코드 작성

- 최근 1일, 1달, 3달만 조회됨
  - 실시간 수집 필요
- 특정 항목 추출 곤란
  - 모두 가져와서 후처리 과정 분리
- Data 가 Spreadsheet 형태로 제공되지 않아서 수집후 후처리 필요
- 5km x 5km 해상도 전국을 분할





환경분야 빅데이터 수집사례(2/4)

- (기상자료개방포털) 데이터 > 날씨예보 | 압축파일 다운로드

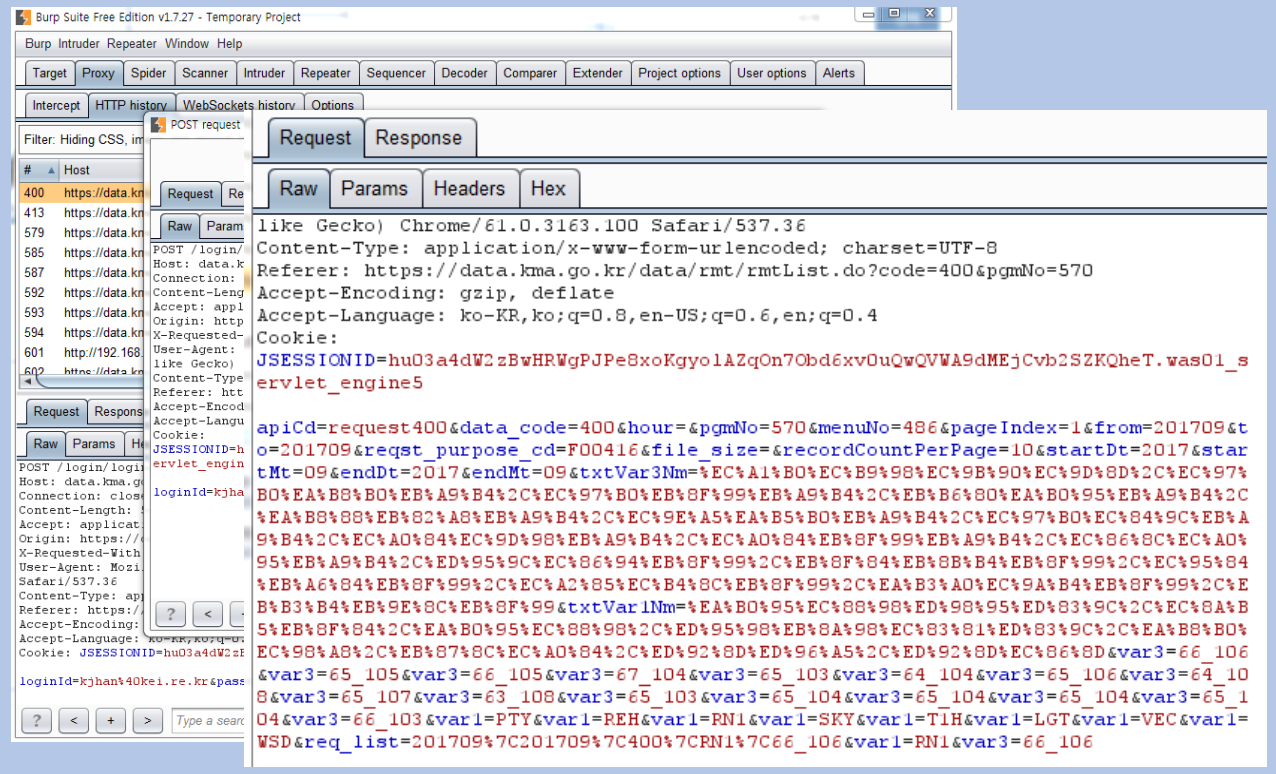
1. 실황분석자료	2. 초단기예보	3. 단기예보	4. ~~예보버전조화~~(없음)

데이터 식별 / 데이터 서비스 탐색

웹페이지 분석(전문SW 필요)

반자동화 코드 작성 + 전문SW 병행

- 로그인해야만 압축 파일 데이터 다운로드 가능
- HTML code 분해 보다는 전문 SW 사용 필요
  - 다운로드 과정을 code화 할 수 없음
- 5km x 5km 해상도 전국을 분할



환경분야 빅데이터 수집사례(3/4)

- (공데 포털) 한국환경공단 대기오염정보 조회 서비스 | 오픈API(XML)

1. 측정소별 실시간 측정정보	2. 통합대기환경지수 나쁨이상	3. 시도별 실시간 측정정보
4. 미세먼지/오전 예보통보	5. 시도별 실시간 평균정보	6. 시군구별 실시간 평균정보

데이터 식별 / 데이터 서비스 탐색

메타데이터 확인 / 오픈API 분석

자동화 코드 작성

```
list_station = []

for station in contents_html.find_all('item'):
    if str(station.find('stationname')) != 'None':
        list_station.append({"stationname": station.find('stationname').text,
                             "addr": station.find('addr').text,
                             "year": station.find('year').text,
                             "oper": station.find('oper').text,
                             "mangname": station.find('mangname').text,
                             "item": station.find('item').text,
                             "dmx": station.find('dmx').text,
                             "dmy": station.find('dmy').text})

cols = ["stationname","addr","year","oper","mangname","item","dmx","dmy"]
#cols = ["측정소명","주소","설치년도","관리기관","측정망","측정항목","위도","경도"]
df_station = pd.DataFrame(list_station, columns=cols)

df_station
```

Out[15]:

	stationname	addr	year	oper	mangname	item	dmx	dmy
0	반송로	경남 창원시 의창구 원이대로 450(시설관리공단 실내수영장 앞)	2008	경상남도보건환경연구원	도로변대기	SO2, CO, O3, NO2, PM10	35.232222	128.671389
1	사파동	경남 창원시 성산구 창이대로 706번길 16-23(사파민원센터)	2009	경상남도보건환경연구원	도시대기	SO2, CO, O3, NO2, PM10	35.221729	128.69825
2	경화동	경남 창원시 진해구 경화로16번길 31(병암동주민센터)	1994	경상남도보건환경연구원	도시대기	SO2, CO, O3, NO2, PM10, PM2.5	35.154972	128.689578

- 특정 항목 추출 곤란
  - 모두 가져와서 후처리 과정 분리
- Data가 Spreadsheet 방식으로 정리되어 있지 않아서 수집 후 후처리 필요

환경분야 빅데이터 수집사례(4/4)

- 에어코리아 | 웹(HTML)

1. (실시간)우리동네대기 정보	2. (실시간)시도별 대기정보	3. (실시간)미세먼지 정보
4. 측정소별 확정자료	5. 측정망·항목별 확정자료	6. 연월보 / 최종확정자료

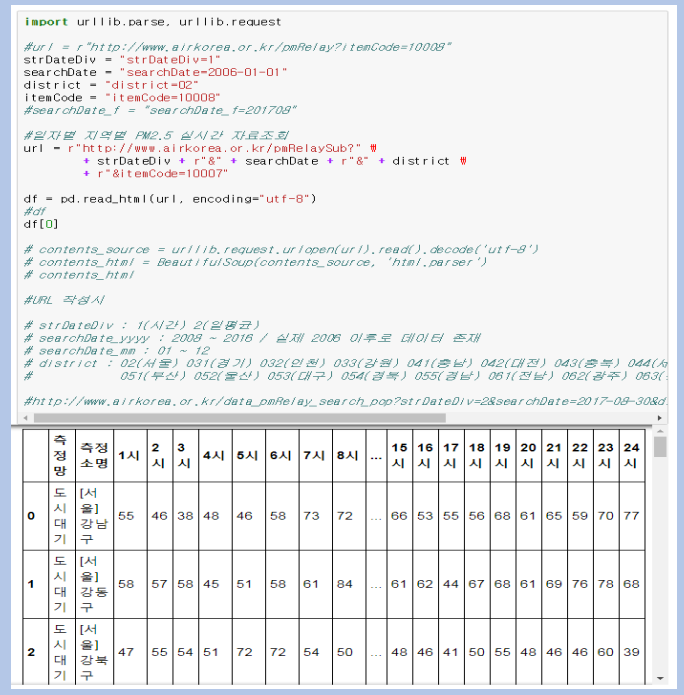
데이터 식별 / 데이터 서비스 탐색

웹페이지 분석

자동화 코드 작성

- HTML 분석만하면 바로 저장 가능함

측정망	측정소명	1시	2시	3시	4시	5시	6시	7시	8시	9시	10시
도시대기	[서울]강남구	32	31	31	30	31	32	33	33	32	30
도시대기	[서울]강동구	38	41	40	39	38	31	36	36	38	42
도시대기	[서울]강북구	20	20	19	19	19	18	18	18	24	33
도시대기	[서울]강서구	43	49	47	53	45	49	48	67	73	83
도시대기	[서울]관악구	33	30	32	31	29	31	30	37	34	36
도시대기	[서울]광진구	31	29	30	28	29	30	32	32	30	30
도시대기	[서울]구로구	33	34	34	32	32	30	35	37	39	36
도시대기	[서울]금천구	35	33	32	32	29	29	29	31	34	37
도시대기	[서울]노원구	36	32	31	32	31	30	32	35	34	37
도시대기	[서울]도봉구	35	36	38	30	36	29	32	35	38	42
도시대기	[서울]동대문구	32	35	33	38	33	36	33	33	34	41
도시대기	[서울]동작구	39	39	38	36	33	34	35	38	45	46
도시대기	[서울]마포구	27	25	26	25	25	25	23	30	32	36
도시대기	[서울]서대문구	32	31	28	28	34	29	26	30	31	26
도시대기	[서울]서초구	34	37	37	30	31	32	33	32	37	34
도시대기	[서울]성동구	34	36	36	32	35	35	35	38	45	47



```
import urllib.parse, urllib.request

#url = r"http://www.airkorea.or.kr/pmRelay?itemCode=10008"
strDateDiv = "strDateDiv=1"
searchDate = "searchDate=2006-01-01"
district = "district=02"
itemCode = "itemCode=10008"
#searchDate_f = "searchDate_f=201708"

#일자별 지역별 PM2.5 실시간 자료조회
url = r"http://www.airkorea.or.kr/pmRelaySub?" ₩
        + strDateDiv + r"&" + searchDate + r"&" + district ₩
        + r"&itemCode=10007"

df = pd.read_html(url, encoding="utf-8")
#df
df[0]

# contents_source = urllib.request.urlopen(url).read().decode('utf-8')
# contents_html = BeautifulSoup(contents_source, 'html.parser')
# contents_html

#URL 작성시

# strDateDiv : 1(시간) 2(일평균)
# searchDate_yyyy : 2009 ~ 2016 / 실제 2006 이후로 데이터 존재
# searchDate_mm : 01 ~ 12
# district : 02(서울) 031(경기) 032(인천) 033(강원) 041(충남) 042(대전) 043(충북) 044(
#            051(부산) 052(울산) 053(대구) 054(경북) 055(경남) 061(전남) 062(광주) 063(

#http://www.airkorea.or.kr/data_pmRelay_search_pop?strDateDiv=2&searchDate=2017-09-30&
```

	측정망	측정소명	1시	2시	3시	4시	5시	6시	7시	8시	...	15시	16시	17시	18시	19시	20시	21시	22시	23시	24시
0	도시대기	[서울]강남구	55	46	38	48	46	58	73	72	...	66	53	55	56	68	61	65	59	70	77
1	도시대기	[서울]강동구	58	57	58	45	51	58	61	84	...	61	62	44	67	68	61	69	76	78	68
2	도시대기	[서울]강북구	47	55	54	51	72	72	54	50	...	48	46	41	50	55	48	46	46	60	39

4. 결론 및 시사점

2017년: 예측/ 연구주제 파악 가능성 확인

- 수치 데이터 예측 알고리듬 3개, 텍스트 데이터 연구동향 분석 알고리듬 3개, 환경 데이터 수집 알고리듬 3개 구축
- 예측 알고리듬: 기존 연구방법론 대비 예측오차 개선
  - LSTM, kNN 공간순환신경망(이동현) : 측정소-시간 미세먼지 오염도 예측오차 10% 개선
  - 심층신경망 (강선아): 시군구-월 장감염 발생빈도 예측오차 25% 개선
  - 랜덤포레스트/Boosting (김진형): 시군구-월 미세먼지 오염도 예측오차 37%/46% 개선
- 반 사실적 실험: 주요 변인의 양적 영향 파악
  - 장감염: 기후변수
  - 미세먼지: 기후변수, 중국대기오염, 2차미세먼지
- 연구동향 분석 알고리듬: 민간 환경 관심과 KEI 연구보고서 동향 추이 비교
  → 새로운 연구주제 도출
  - 새로운 토픽 : 유전자 변형-소음, 보건-데이터 연구
  - 기존 토픽 연구 방향 : 기후변화 총론 연구 → 태풍, 한파. 대설 등 세부 현상 연구

빅데이터 연구 결과 환경정책 활용 가능성

- 1. 정확한 환경오염 단기 예측치 제공 (이동현)
  - 민간의 관심 충족 및 민간 환경 정보 서비스 인프라 제공
- 2. 단기 정책재원 운용 효율성 제고 (강선아, 김진형)
  - 환경위험 예측치가 높은 지역에 우선적으로 환경정책 운용 인력/재원 배치
- 3. 소규모 지역 단위 정책 평가 지표 활용
  - 반 사실적 실험을 이용하여 정책지표가 정책목표의 변화에 미친 영향 정량화
- 4. 기존 정책이 포괄하지 못하고 있는 정책 대상 파악 (김진형)
  - 반 사실적 실험: 기존 정책 대상은 아니지만 정책목표에 큰 변화를 주는 변인 파악
- 1~4 신규 자료 이용 주기적 update 가능

5. 향후 계획

부록 집필 계획

- ‘공간시계열 분석: 시계열 분석과 기계학습’
  - 전통적인 시계열 분석 방법과 기계학습 비교
- ‘텍스트 마이닝: LDA 와 Word2Vec’
  - 텍스트 마이닝 분석 방법인 LDA 와 Word2Vec의 이론적 소개
- ‘환경빅데이터 플랫폼 구성방안’
  - 환경 빅데이터 분석 인프라인 환경 빅데이터 플랫폼 구축 계획

감사합니다