


| | | | | |
|--|--|---|-----|---------------------------------|
| 사 례 명 | 통계분석 및 기계학습 적용을 위한 전국 시도별, 연도별 화학물질 배출량 데이터셋 구성 및 통계와 GIS 기법을 통한 검증 | | | |
| 인적사항 | 기관명 | 한국지질자원연구원 | 부서명 | 지질자원데이터센터 |
| | 직 급 | 석사과정연구생 | 성 명 | 박강현 |
| | 연락처 | ☎ 033-642-6392 (휴대 폰: 010-8806-6392) | | Email ganghyun6392@naver.com |
| | 비 고 | | | |
| 우수사례 요약설명 | <p>환경 빅데이터플랫폼에 존재하는 ‘그린에코스’의 ‘화학물질 배출량 정보’를 기반으로, 빅데이터를 이용한 데이터 분석(통계분석) 및 인공지능(머신러닝, 딥러닝) 융합에 적합한 형태로 데이터를 재구성하였다.</p> <p>이후 새로 정립된 데이터를 바탕으로, 데이터 분석가의 교재 및 놀이터로서, 빅데이터를 체계적이고 다양한 방법으로 접근하여 개인의 역량증진과 환경 빅데이터 플랫폼의 빅데이터 구성 양식에 대한 제안을 구상하며 제작되었다.</p> <p>본 작업에서는 기존의 데이터셋을 통계, 기계학습에 적합한 데이터셋으로 재구성하여, 다양한 기법 적용에 용이한 형태로 전처리를 진행하였고, 통계분석, GIS를 통한 공간적 시각화를 시도하였다.</p> | | | |
| <p>위와 같이 「2020년 환경 빅데이터 활용 우수사례 공모전」에 신청서를 제출합니다.</p> <p>붙 임 : 환경 빅데이터 활용 우수사례 보고서 1부.</p> <p style="text-align: center;">2020. 10 . 15 .</p> <p style="text-align: right;">신청자 <u>박강현</u> (서명 또는 인)</p> <p style="text-align: right;">한국수자원공사장 귀하</p> <p style="text-align: center;"></p> | | | | |

환경 빅데이터 활용 우수사례 보고서

사례명

통계분석 및 기계학습 적용을 위한 전국 시도별, 연도별 화학물질 배출량 데이터셋 구성 및 통계와 GIS 기법을 통한 검증

□ 추진배경 (HY헤드라인M 16, 줄간격 160, 문단아래 15)

○ 데이터 분석가의 실제적인 데이터 운용을 통한 실무 기술 연마

- 데이터 분석, 관리를 공부하는 학생으로서 이론적인 지식 외에도 실무적인 기술의 필요성을 느끼고, 환경 빅데이터 플랫폼에서 제공하는 다양한 데이터들을 직접적으로 다루며, 실제 존재하는 데이터셋의 구조와 한계를 확인하고, 기존의 데이터셋을 통계와 GIS, 기계학습에 적용하기 위한 적합 데이터셋으로 전처리하는 과정을 진행하고, 그 데이터셋을 기반으로 실질적인 기술을 적용해봄으로써, 개인 역량 증진의 기회로 사용하고자 하였다.

□ 추진과정

○ 데이터 수집과정 및 방향 설정

['year', 'sn_no', 'company', 'CTP', 'SGG', 'address', 'discharge', 'dsc_unit', 'reclaim', 'recl_unit', 'transport', 'trans_unit', 'origin']

| | year | sn_no | company | CTP | SGG | address | discharge | dsc_unit | reclaim | recl_unit | transport | trans_unit | origin |
|-------|------|-------|------------------|------|---------|-------------------------------|-----------|----------|---------|-----------|-----------|------------|-------------------|
| 0 | 2001 | 1 | (유)남해환경 | 전라남도 | 무안군 | 전라남도 무안군 삼향중앙로 140-51 (삼향읍) | 0 | kg/yr | 0 | kg/yr | 43,896 | kg/yr | 화학물질종량정보시스템(ICIS) |
| 1 | 2001 | 2 | (유)탑코리아 | 경기도 | 화성시 | 경기도 화성시 장안공단8길 42 (장안면) | 0 | kg/yr | 0 | kg/yr | 10,497 | kg/yr | 화학물질종량정보시스템(ICIS) |
| 2 | 2001 | 3 | (유)보금 | 경상남도 | 양산시 | 경상남도 양산시 소주공단5길 3 (주남동) | 40 | kg/yr | 0 | kg/yr | 13,803 | kg/yr | 화학물질종량정보시스템(ICIS) |
| 3 | 2001 | 4 | (유)삼승 | 경상남도 | 장원시 성안구 | 경상남도 장원시 성안구 정동로62번길 30 (성주동) | 27,400 | kg/yr | 0 | kg/yr | 0 | kg/yr | 화학물질종량정보시스템(ICIS) |
| 4 | 2001 | 5 | (유)셀가드코리아 | 충청북도 | 청주시 청원구 | 충청북도 청주시 청원구 연구단지로 208 (오창읍) | 82,929 | kg/yr | 0 | kg/yr | 21 | kg/yr | 화학물질종량정보시스템(ICIS) |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 48391 | 2017 | 3794 | 희성플러머(주) | 충청남도 | 천안시 서북구 | 충청남도 천안시 서북구 천안대로 2131 (성환읍) | 186,927 | kg/yr | 0 | kg/yr | 58,381 | kg/yr | 화학물질종량정보시스템(ICIS) |
| 48392 | 2017 | 3795 | 희성파엘(주) | 충청남도 | 당진시 | 충청남도 당진시 부곡공단4길 28-76 (송악읍) | 15 | kg/yr | 0 | kg/yr | 523,766 | kg/yr | 화학물질종량정보시스템(ICIS) |
| 48393 | 2017 | 3796 | 히트세코리아(주) | 경기도 | 시흥시 | 경기도 시흥시 공단1대로 143 (정왕동) | 0 | kg/yr | 0 | kg/yr | 292,315 | kg/yr | 화학물질종량정보시스템(ICIS) |
| 48394 | 2017 | 3797 | 히트세코리아(주) AMC 센터 | 경기도 | 시흥시 | 경기도 시흥시 희망공원로 250 (정왕동) | 15 | kg/yr | 0 | kg/yr | 133,141 | kg/yr | 화학물질종량정보시스템(ICIS) |
| 48395 | 2017 | 3798 | 히타치금속한국 주식회사 | 경기도 | 평택시 | 경기도 평택시 한산길 64 (정북읍) | 0 | kg/yr | 0 | kg/yr | 0 | kg/yr | 화학물질종량정보시스템(ICIS) |

48396 rows x 13 columns

그림 1 - 화학물질 배출량 정보 데이터프레임

○ 데이터 선정단계

환경 빅데이터 플랫폼의 「그린 에코스」에서 제공하는 「화학물질 배출량 정보」 데이터(CSV 형식)를 다운로드하였다. 이후 파이썬 Pandas 라이브러리를 통해, CSV데이터를 데이터프레임으로 구성하고, 컬럼명을 가시적이게 재구성한다.

해당 데이터셋의 경우, 연도단위의 시계열로 구성되어 있으며, 각각 데이터의 공간적 주소가 명시되어 있다. 또한 결측행이 적으며, 데이터는 13개의 열과 48396개의 행으로 구성되어, 데이터 분석 과정에서 신뢰성과 다양성을 충족할 수 있다고 판단하여 선정하였다.

○ 목적성 설계 단계

기존 데이터에 통계 외에도 GIS를 통한 공간적인 분석을 동반 할 수 있도록, 시군구 코드가 담긴 CSV파일을 가져와 기존 데이터프레임에 삽입하고, 분석상 불필요한 부분과 결측값이 존재하는 행을 제거하여, 데이터 프레임을 단순화한다.

2001년 부터 2017년까지의 데이터 개수는 1022, 1197, 1383, 2878, 2741, 2769, 3009, 2945, 2917, 2985, 3159, 3268, 3435, 3524, 3634, 3732, 3798개

그림 2 - 본래 데이터의 구성 개수 정보

○ 데이터 이해 단계

Raw 데이터를 구성하는 데이터의 확인 및 이해과정으로, 2001년부터 근년에 가까워질수록, 한해마다 화학물질을 배출하는 업체 혹은 기관이 지속적으로 증가함을 알 수 있다.

□ 추진내용

○ 분석을 위한 데이터셋 편집 및 중합 코드 작성 및 데이터셋 분할 구성 작업

| | year | SGG | SGG_code | CTP | CTP_code_x | discharge | reclaim | transport | CTP_dis_sum | CTP_rec_sum |
|-----|------|---------|----------|------|------------|-----------|---------|-----------|-------------|-------------|
| 909 | 2001 | 홍천군 | 42720 | 강원도 | 42 | 0 | 0 | 0 | 9620.0 | 0.0 |
| 22 | 2001 | 원주시 | 42130 | 강원도 | 42 | 3,369 | 0 | 0 | 9620.0 | 0.0 |
| 619 | 2001 | 원주시 | 42130 | 강원도 | 42 | 211 | 0 | 0 | 9620.0 | 0.0 |
| 779 | 2001 | 원주시 | 42130 | 강원도 | 42 | 6,040 | 0 | 226,123 | 9620.0 | 0.0 |
| 666 | 2001 | 시흥시 | 41390 | 경기도 | 41 | 0 | 0 | 0 | 2883999.0 | 19464.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 257 | 2001 | 괴산군 | 43760 | 충청북도 | 43 | 7,505 | 0 | 185,695 | 2264983.0 | 1257.0 |
| 470 | 2001 | 청주시 청원구 | 43114 | 충청북도 | 43 | 1,331 | 0 | 0 | 2264983.0 | 1257.0 |
| 711 | 2001 | 청주시 청원구 | 43114 | 충청북도 | 43 | 67 | 0 | 0 | 2264983.0 | 1257.0 |
| 83 | 2001 | 괴산군 | 43760 | 충청북도 | 43 | 3,871 | 0 | 4,716 | 2264983.0 | 1257.0 |
| 914 | 2001 | 청주시 흥덕구 | 43113 | 충청북도 | 43 | 2 | 0 | 0 | 2264983.0 | 1257.0 |

1022 rows × 10 columns

그림 3 - 연도에 따른 시도별 배출량과 자가매립량 데이터셋(2001년)

- 해당 데이터셋은 각각 연도별로 데이터셋을 나눈 후, 시도별 배출량과 자가매립량을 계산하여 하나의 열로 삽입한 결과이다.

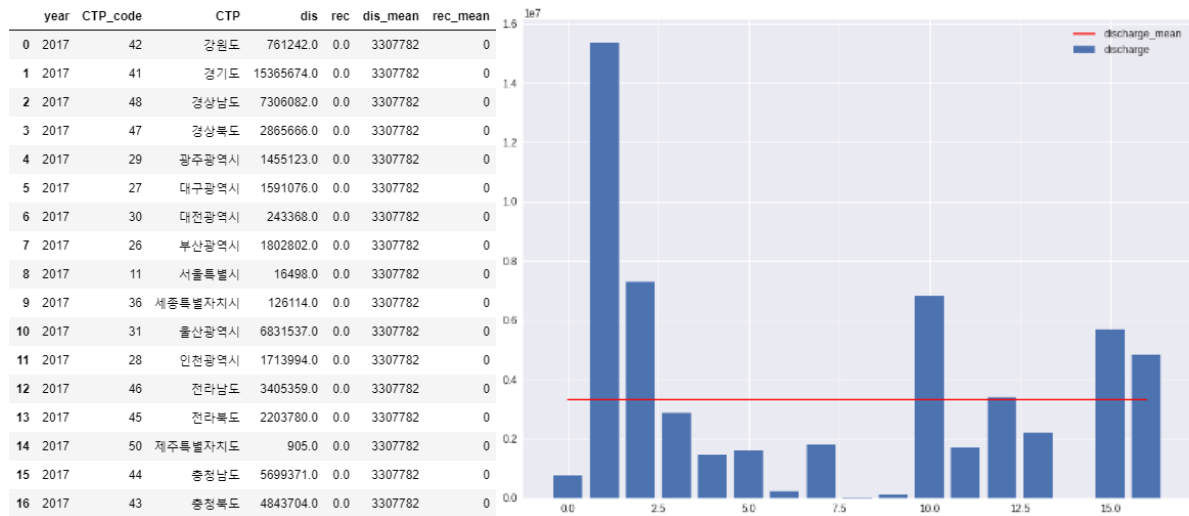


그림 4.5 - 2017년 기준 시도별 화학물질 배출량 데이터셋

- 2017년 기준의 시도별 화학물질 배출량을 나타낸다. 경상남도, 울산광역시, 충청남도, 충청북도가 전국 평균을 상회하며 경기도가 가장 큰 비중을 차지하는 것을 볼 수 있다.

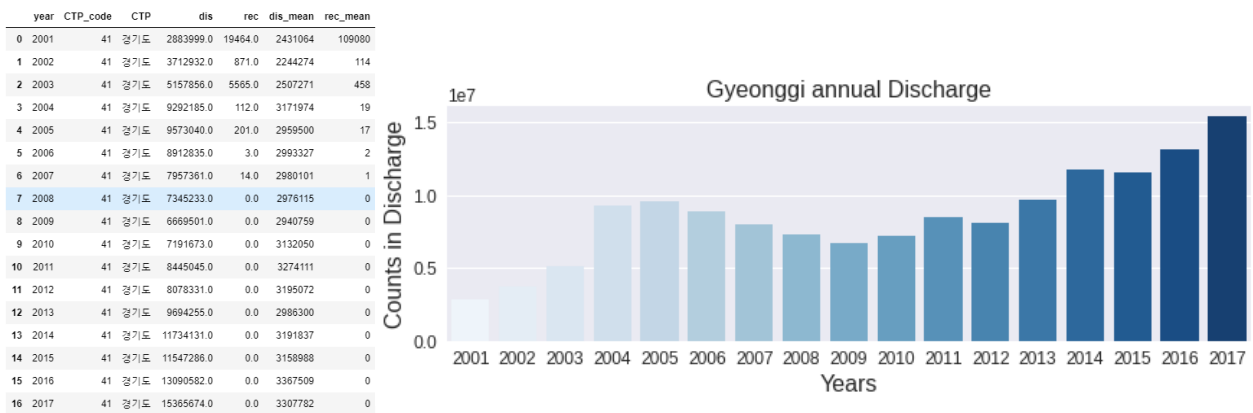


그림 6.7 - 경기도 지역 연도별 화학물질 배출량 데이터셋

- 다 지역 대비 화학물질 배출량이 가장 많은 경기도의 연도별 배출량을 나타낸 그래프로, 2000년대 후반, 잠시 줄어드는 추세에서 다시 지속적으로 배출량이 늘어남을 볼 수 있다.

○ 구성된 데이터셋을 이용한 통계 분석과 공간적 분석

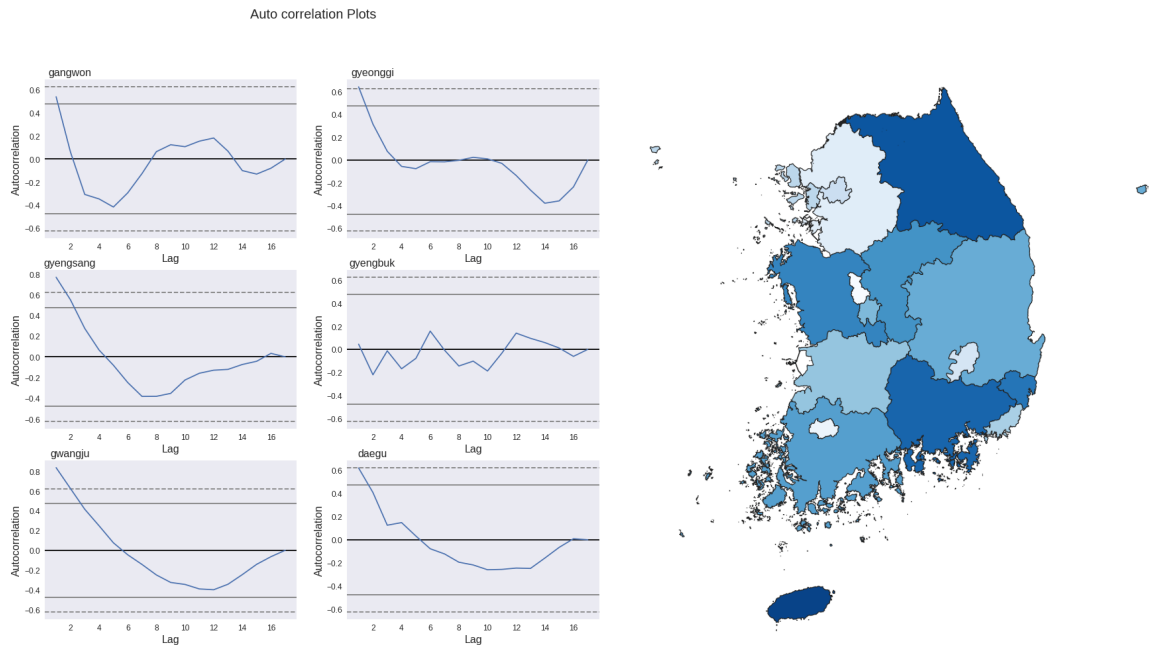


그림 8,9 - 구성된 데이터셋을 이용한 자기상관성 그래프와 GIS 시각화

- 그림 6에서 구성된 데이터셋을 바탕으로 시도별 연도별 자기상관성 분석을 진행하였으며(그림8), 목적성 설계단계에서 삽입한 시도별 code를 통하여, 전국 화학물질 배출량을 시각화하였다.

□ 추진성과

○ 데이터셋 처리방식의 기준 정립

- 빅데이터 플랫폼에 존재하는 다양한 데이터를 직접적으로 다루며, 데이터 구성방식에 대한 정리 체계를 구성하였다.

- 1 - 목적에 따른 필요한 정보를 추가하기(GIS 시각화를 위한 공간정보 삽입)
- 2 - 데이터 분석가의 가시성을 위한 데이터셋 정리제거 작업
- 3 - 기능적 활동(통계분석)을 위한 기존 데이터 기반의 새로운 데이터 생성(업체 및 기관별 화학물질 배출량, 자가매립량의 시도별 합산 컬럼의 생성과정)
- 4 - 효율 및 차후 다양한 조합의 기법 시도를 위한 동일 데이터셋의 다양한 구성 조립(통합 정제 데이터셋, 연도에 따른 시도별 데이터셋, 시도별 연도별 데이터셋)
- 5 - 통계적, 공간적 분석

○ 데이터 분석가의 역량증진

- 환경 빅데이터 플랫폼은 환경 데이터과학자를 희망하는 필자에게, 역량 증진을 위한 가장 적합한 플랫폼으로, 습득한 이론적인 지식을 실제 데이터에 대한 적용을 통해 실제적인 기술로 탈바꿈 할 수 있는 공간이자 놀이터이다.

데이터를 다루는 분석가들에게 그들이 연마해왔던 전처리, 데이터 정리, 통계적 분석과 GIS기법의 적용, 나아가 기계학습 기술을 집합한 데이터 분석기술을 통해 방대한 데이터를 새로운 정보로 탈바꿈 시킬 수 있도록 도와주는 또하나의 학교라고 생각한다.

<작 성 요 령>

- 서식 변경 가능, 위의 내용이 포함되게 **5페이지 이내 개조식**으로 작성
(※ 우수사례 보고서 서식 외 추가 내용이 있는 경우, 별첨으로 작성 가능)
- 글꼴준수, 용지여백은 위·아래 15, 머리·꼬리말 15, 좌·우 20