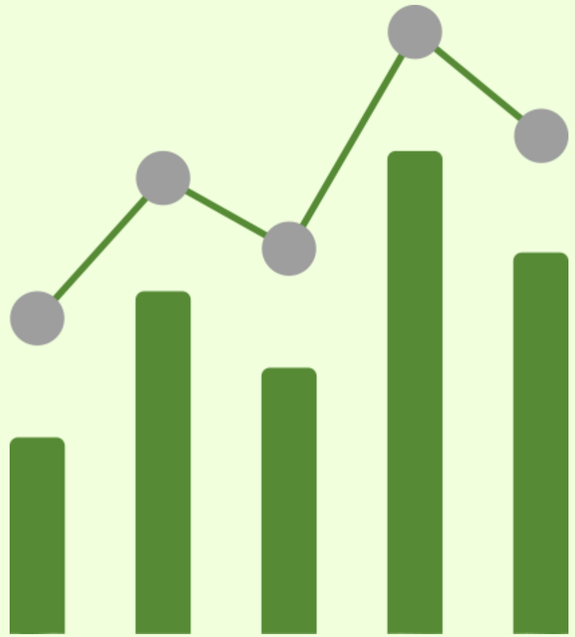
제1회 2023년 지역 치안 안전 데이터 분석 공모전

충남·세종·대전 지역 교통사고 분석 및 예측

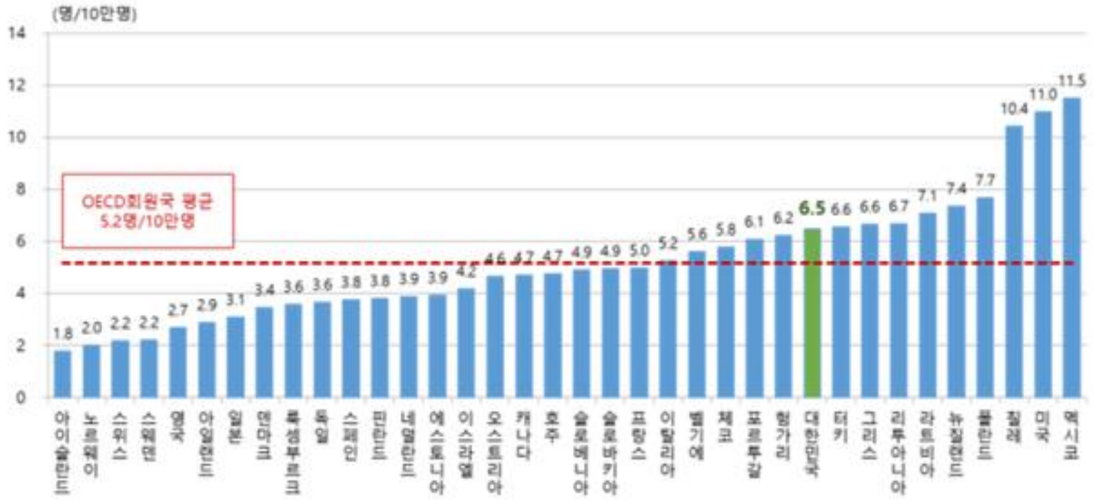
HDBSCAN을 활용한 교통사고 Hotspot 분석 및 예측



강양이 방범대
강훈, 양정열, 이승준, 이채원

BACKGROUND

분석 배경 및 필요성



OECD 회원국 인구 10만명 당 교통사고 사망자 비교(2019년 기준) / 표 = 도로교통공단 제공

"한국, 교통사고 사망자수 OECD 36개국 중 27위... 10만명당 6.5명"

우리나라 교통사고 관련 각종 통계 지수가 후진국 수준을 벗어나지 못한 것으로 드러났다. 도로교통공단이 국가별 교통사고 현황을 분석한 'OECD 회원국 교통사고 비교 보고서(2021년판)'에 따르면 인구 10만명 당 교통사고 사망자 27위, 자동차 1만대 당 사망자 31위, 특히 교통사고 사망자 중 보행자 비율이 38.9%로 OECD 회원국 가운데 최하위를 기록했다.

출처 : 안전신문(https://www.safetynews.co.kr)

현대사회에서 교통사고는 중요한 사회적 문제이다. 한국의 교통사고 건수는 지속적으로 감소하고 있지만, 교통사고 사망자 수는 여전히 OECD 국가들에 비해 높은 수준을 유지하고 있으며 국내 교통안전수준 또한 하위권에 머물러 있다. 교통사고 관련 데이터는 수집·분석을 통해 교통사고 위험을 미리 예측하고 그에 맞는 대책을 세울 수 있기 때문에 교통사고를 예방하는 데 중요한 요소이다. 본 프로젝트에서는 스마트 치안 빅데이터 플랫폼에서 제공한 데이터를 활용하여 충남, 세종, 대전의 교통사고 발생 현황을 분석하고 교통사고를 선제적으로 예측할 수 있는 모델을 생성한다.

PURPOSE

분석/시각화 목적

교통사고는 다양한 요인으로 인해 일부 지역에서 집중되는 경향이 있다. 도로안전을 보장하고 교통사고를 줄이기 위해서는 교통사고가 많이 발생하는 지리적 위치를 파악하는 것이 필수적이다. 본 프로젝트에서는 교통사고와 관련된 데이터를 수집하여 교통사고 Hotspot을 분석하고, 주요 Hotspot의 교통사고 발생을 예측한다.

교통사고 Hotspot 분석

- 데이터 군집화를 통해 교통사고 발생 밀도가 높은 Hotspot 영역 확인
- 각 Hotspot의 시간적, 계절적 요인에 따른 교통사고 발생 패턴 분석

Hotspot 교통사고 발생 예측

- 각 Hotspot별로 교통사고 발생 예측 모델을 생성하여 교통사고 발생 여부를 예측

교통사고 Hotspot 분석

교통사고 Hotspot 분석 로드맵

HDBSCAN을 활용하여
사고 발생지점의 위도·경도 좌표 데이터를 군집화,
Spotfire를 활용하여 지역별 교통사고가
밀집되어 있는 Hotspot 확인

지역별 Hotspot clustering

 # 01

TIBCO Spotfire Partner

 # 02

지역별 Hotspot Heatmap

Folium

Folium을 활용하여
지역별 교통사고의 밀도를
Heatmap으로 확인

선정된 주요 Hotspot의 영역을 정의하기 위해
QGIS를 활용하여 50m x 50m Grid 생성,
연도 / 계절 / 요일 / 밤낮 요인별로 분석

지역, 요인별 주요 Hotspot Grid

 # 03

QGIS

 # 04

Time series clustering

시계열 간 유사도를 바탕으로 생성한
시계열 데이터 군집을
Grid를 활용하여 분석

데이터 전처리

(KP2020, KP2021)

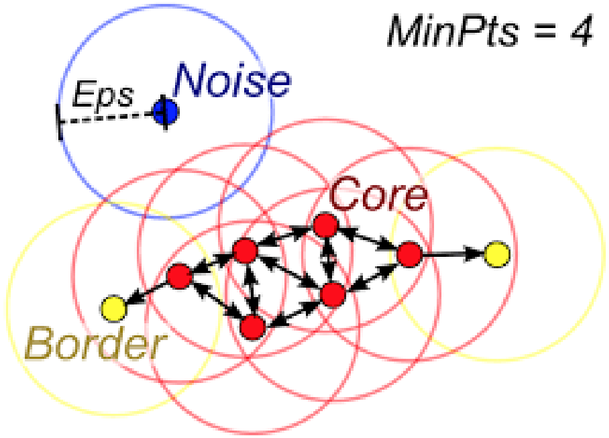
변수명	변수설명
RECV_DEPT_NM	접수부서 코드
RECV_CPLT_DM	접수완료일시
NPA_CL	경찰청구분
EVT_STAT_CD	사건상태코드
EVT_CL_CD	사건종별코드
RPTER_SEX	신고 성별
HPPN_PNU_ADDR	발생지점(PNU)
HPPN_X	발생좌표X
HPPN_Y	발생좌표Y
SME_EVT_YN	동일사건여부

(NPA2020)

변수명	변수설명
RECV_CPLT_DT	접수완료일자
RECV_CPLT_TM	접수완료시간
NPA_CL	경찰청구분
EVT_STAT_CD	사건상태코드
EVT_CL_CD	사건종별코드
RPTER_SEX	신고 성별
HPPN_OLD_ADDR	발생구주소
HPPN_X	발생좌표X
HPPN_Y	발생좌표Y
SME_EVT_YN	동일사건여부

- 교통사고 발생 밀도가 높은 Hotspot을 선정하기 위하여 제공받은 KP2020, KP2021, NPA2020 데이터셋을 활용
- NPA2020 데이터셋의 접수완료일자(RECV_CPLT_DT)와 접수완료시간(RECV_CPLT_TM) 열을 병합하고 Datetime 형식으로 변환
- KP2020, KP2021 데이터셋의 접수완료일시(RECV_CPLT_DM) 또한 Datetime 형식으로 변환
- 교통사고 데이터만을 활용하기 위하여 사건종별코드(EVT_CL_CD)가 401(교통사고)인 행을 추출한 후 세개의 데이터셋을 병합
- 위도(HPPN_Y), 경도(HPPN_X) 값이 결측치인 행은 제거
- 사고 발생지점 주소에 대한 결측치는 kakaomap API[1]를 활용해 획득한 주소값으로 대체
- 사고 발생지점 주소, 위도, 경도 열을 제외한 분석에 불필요한 나머지 열 제거
- 사고 발생지점 주소를 기준으로 전체 데이터셋을 충남 / 세종 / 대전 데이터셋으로 분할하고 각 데이터셋에 대하여 Hotspot 분석을 진행

HDBSCAN



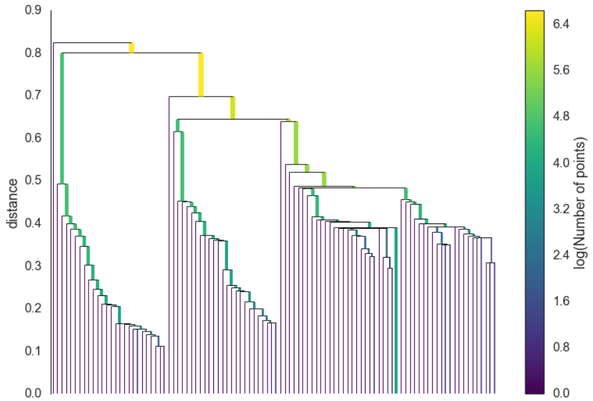
DBSCAN

DBSCAN

- DBSCAN(Density-based spatial clustering of applications with noise)은 밀도 기반의 군집화 방법론으로써 epsilon과 MinPts라는 두가지 파라미터를 가짐
- Epsilon은 이웃으로 간주될 두 점 사이의 최대 거리를 결정
- MinPts는 군집을 형성하는데 필요한 최소 데이터 포인트의 개수를 결정
- 군집에 속하지 못한 데이터 포인트는 노이즈로 간주
- 밀도가 서로 다른 군집을 포함하고 있는 데이터셋에서는 최적의 파라미터들을 찾기 어려울 수 있음(참고문헌[1])

HDBSCAN

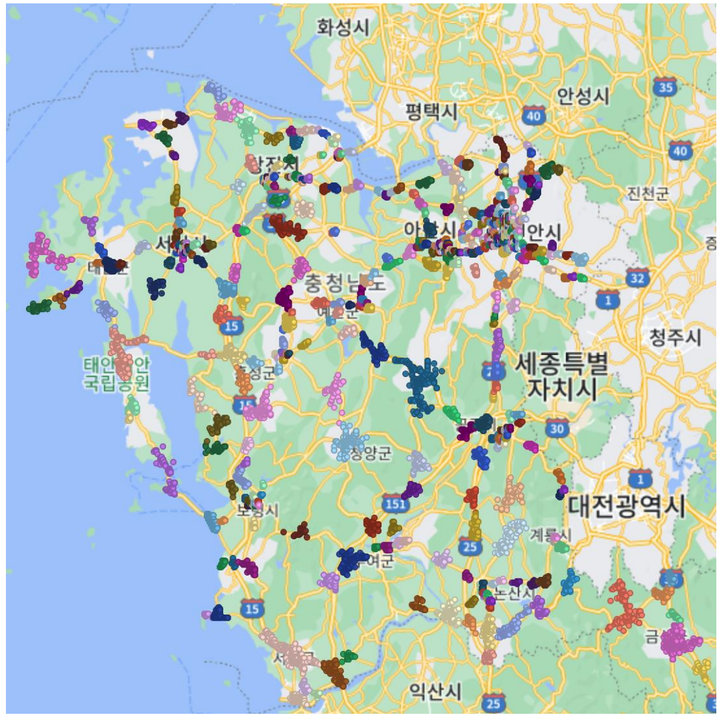
- HDBSCAN(Hierarchical DBSCAN)은 DBSCAN을 계층적 군집화로 확장한 방법론
- HDBSCAN은 Excess of Mass 개념을 활용하여 전체 군집의 안정성을 최대화하는 최적의 epsilon을 산출
- 본 프로젝트에서는 HDBSCAN을 활용하여 교통사고 발생 밀도가 높은 영역을 포착하고 Hotspot으로 선정
- 또한 Excess of Mass 알고리즘을 통해 산출된 최적의 epsilon을 사용하여 교통사고 발생 위도·경도 좌표계를 군집화
- 데이터 포인트 간 거리를 측정하기 위해서 위도·경도 좌표계 사이의 거리를 측정하는 함수인 Haversine distance를 거리척도로 사용



Dendrogram of HDBSCAN

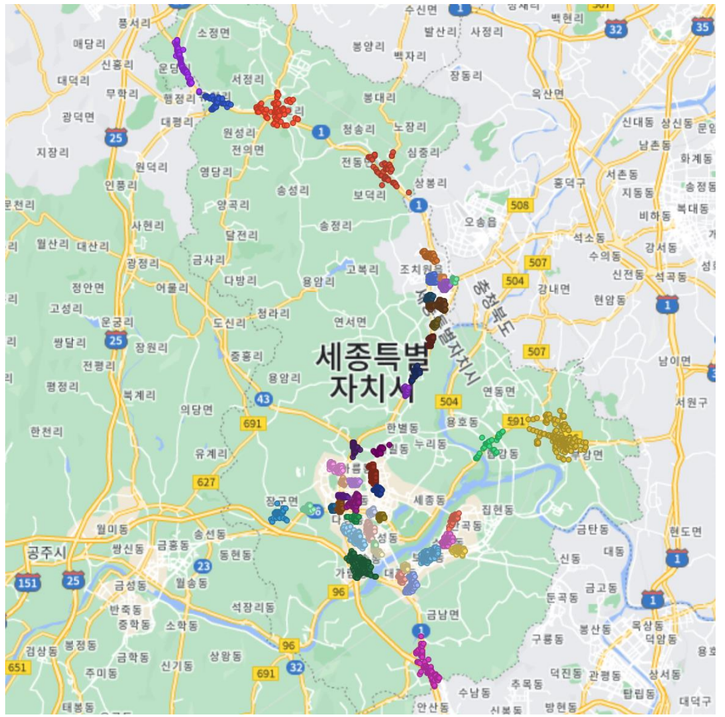
HDBSCAN

충남



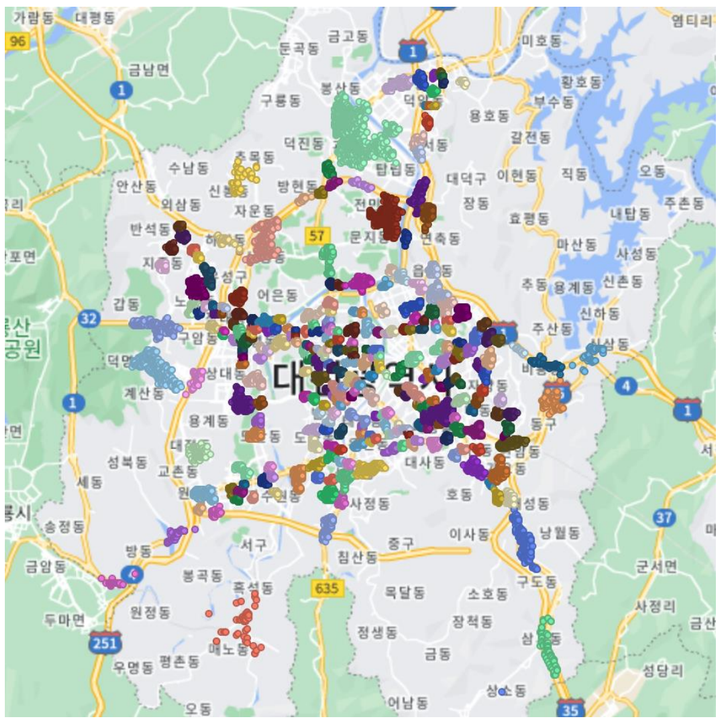
MinPts = 40

세종



MinPts = 50

대전

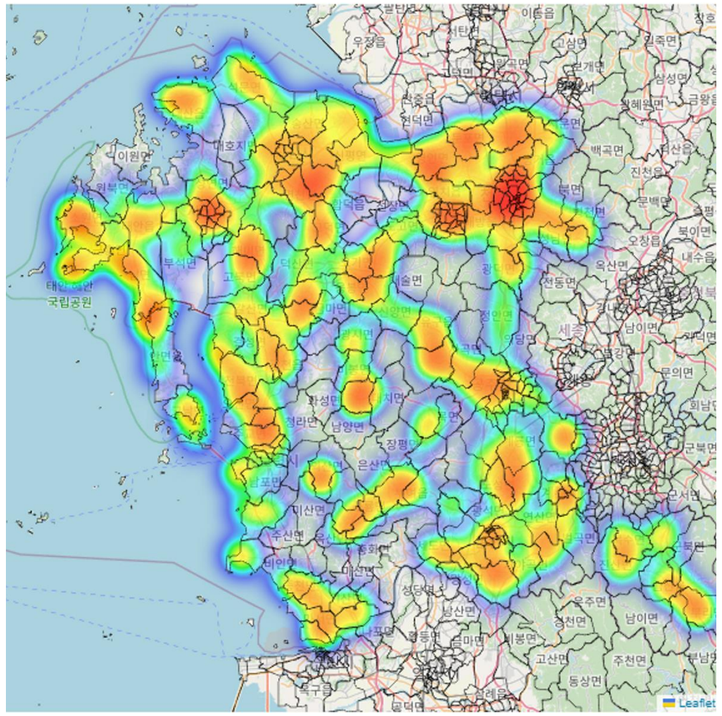


MinPts = 50

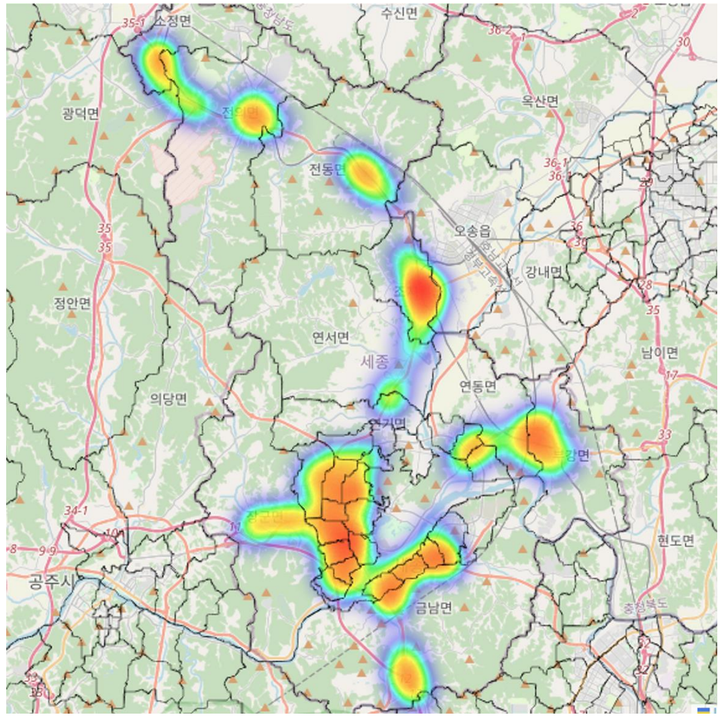
- Spotfire를 활용하여 군집화 된 교통사고 발생 위도·경도 좌표들을 시각화 (노이즈 제거)
- 교통사고가 밀집되게 발생하여 군집화 된 영역을 교통사고 Hotspot 지역으로 선정
- 충남 558개, 세종 46개, 대전 291개의 Hotspot을 확인

Hotspot 시각화 - Heatmap

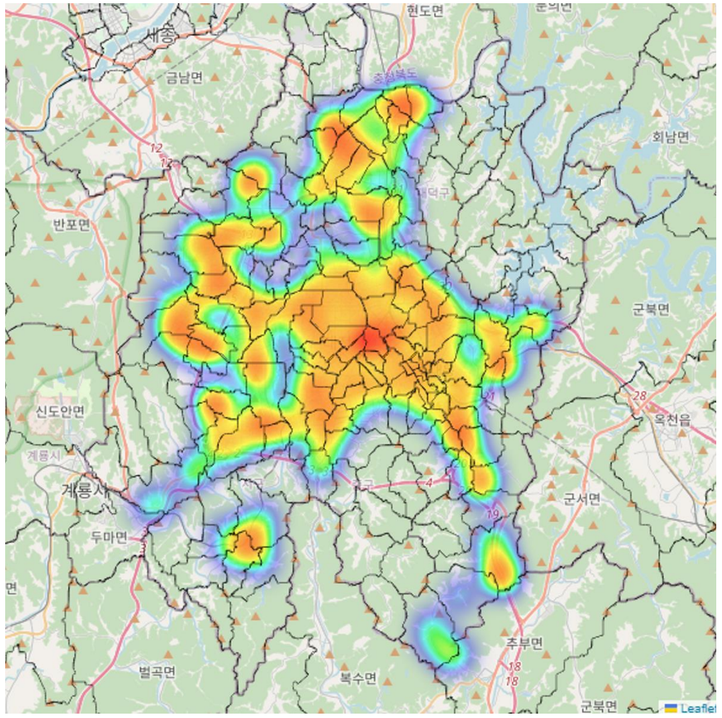
충남



세종



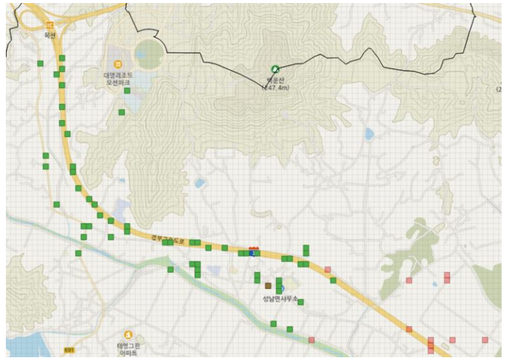
대전



- Folium을 활용하여 충남, 세종, 대전 별 Heatmap 시각화
- Hotspot 데이터의 위도·경도 활용
- 밀도가 높은 영역을 붉은색으로 표시

Hotspot Grid 시각화 - 충남

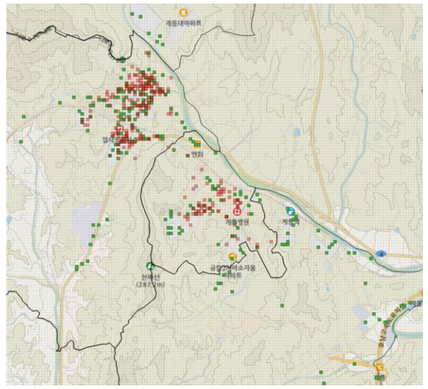
연도별



2020년
2021년, 2022년

○ 성남면 경부 고속도로

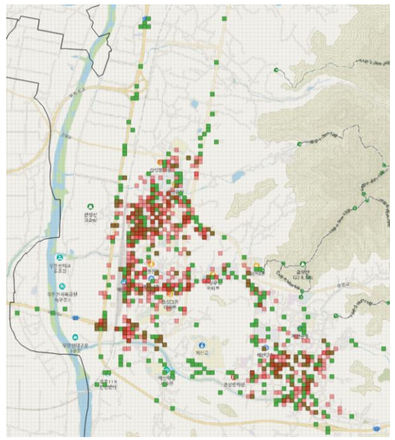
- 2021년, 2022년 :
Hotspot이 존재하지 않음
- **2020년 :**
천안IC와 옥천IC 사이의 고속도로에서 사고 다수 발생



2021년
2020년, 2022년

○ 충청남도 화천면 지곡리

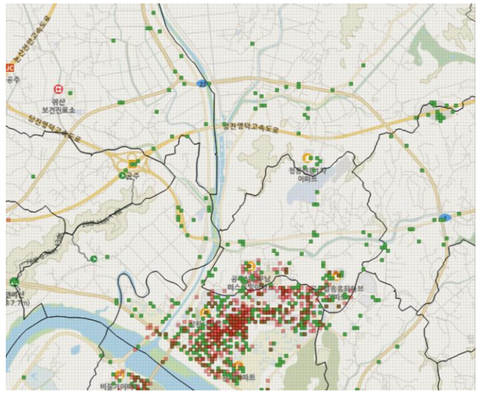
- 2020년, 2022년 :
계룡시청, 계룡병원을 중심으로 사고 밀집
- **2021년 :**
계룡시청, 계룡 병원 뿐만 아니라 주변 도로에도 사고 다수 발생



2021년
2020년, 2022년

○ 예산종합터미널, 예산군청

- 2020년, 2022년 :
예산종합터미널, 예산군청을 중심으로 사고 밀집
- **2021년 :**
예산종합터미널, 예산군청 뿐만 아니라 주변 도로에도 사고 다수 발생



2022년
2020년, 2021년

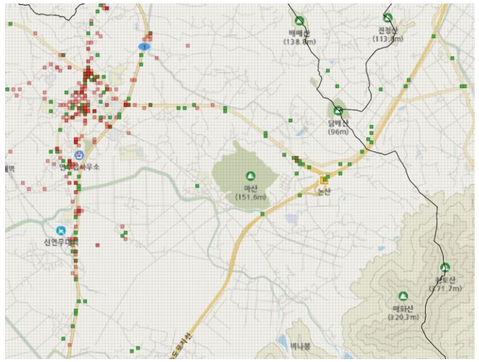
○ 충청남도 공주 종합버스터미널

- 2020년, 2021년 :
종합버스터미널 중심으로 사고 밀집
- **2022년 :**
종합버스터미널 뿐만 아니라 주변 도로에도 사고 다수 발생

Hotspot Grid 시각화 - 충남

* 낮 : 7시 - 19시, 밤 : 19시 – 7시

계절별

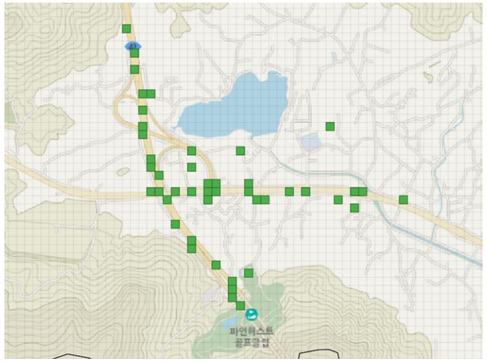


가을
봄, 여름, 겨울

○ 연무대 고속터미널, 호남고속도로

- 가을 :
다른 계절에 비해 연무대 고속터미널, 호남고속도로 방면에서 사고 다수 발생

요일별

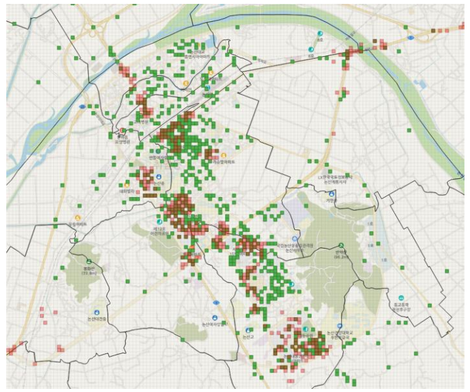


주말
주중

○ 아산시 음봉면 파인허스트골프클럽

- 주중 :
Hotspot이 존재하지 않음
- **주말 :**
골프클럽에서 천안 방면 도로 사고 다수 발생

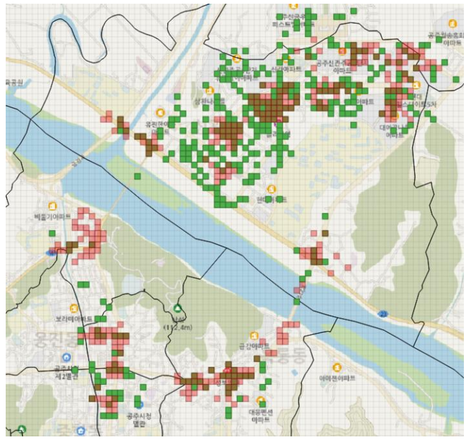
시간별



밤
낮

○ 논산시청 및 논산 버스터미널

- 밤 :
낮에 비해 논상 시청, 논산 버스터미널에서 사고 다수 발생



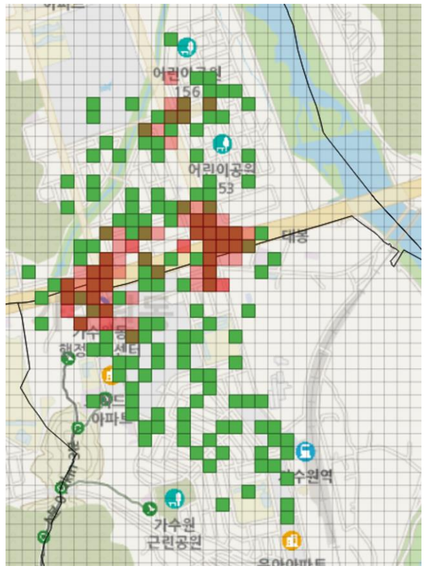
밤
낮

○ 공주 종합버스터미널 및 시내버스터미널

- 밤 :
낮에 비해 공주 종합버스터미널, 시내버스터미널에서 사고 다수 발생

Hotspot Grid 시각화 - 대전

연도별

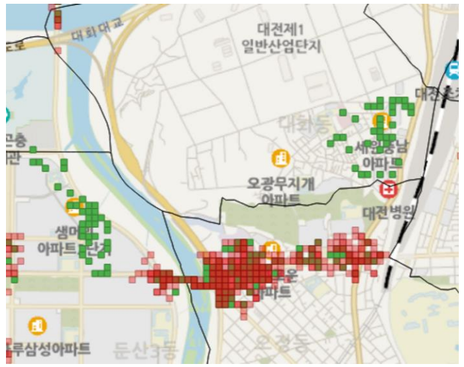


2021년
2020년, 2022년

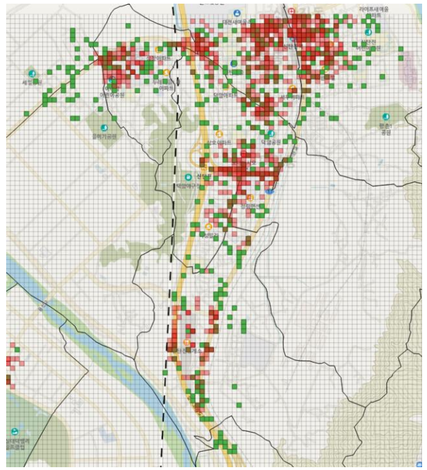
○ 건양대

- 2020년, 2022년 :
건양대 근처 시내의
도로(계백로)주위에 사고 발생
- **2021년 :**
도로(계백로)주위 뿐만 아니라
시내에서도 사고 다수 발생

계절별



가을
봄, 여름, 겨울



겨울
봄, 여름, 가을

○ 대전 곤충생태관, 일반산업단지

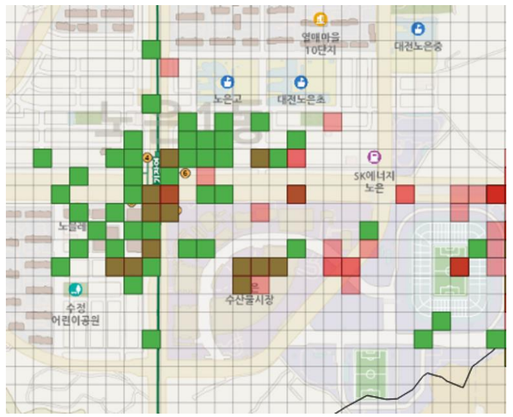
- 가을 :
다른 계절에 비해 곤충생태관,
산업체 출입 방면 길목, 도로에서
사고 다수 발생

○ 을미기공원

- 겨울 :
다른 계절에 비해 신탄진역 주위
도로 시내에서 사고 다수 발생

Hotspot Grid 시각화 - 대전

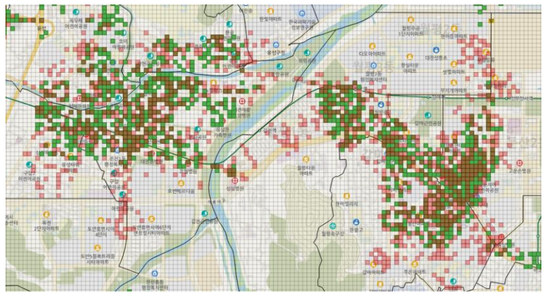
계절별



주말
주중

○ 월드컵경기장역

- 주말 :
주중에 비해 월드컵경기장역 근처 시내 방면에서 사고 다수 발생

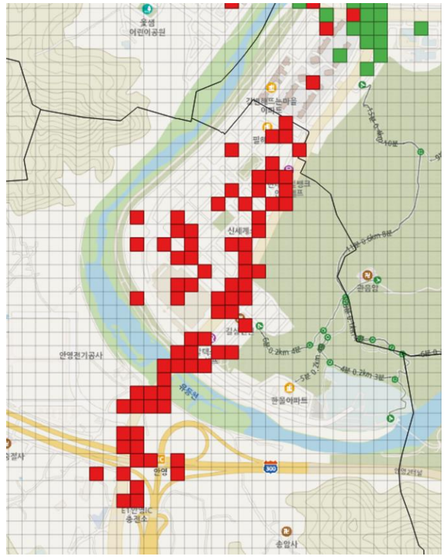


주말
목요일

○ 유성온천역

- 목요일 :
다른 요일에 비해 대전 시내에서 사고 다수 발생

시간별



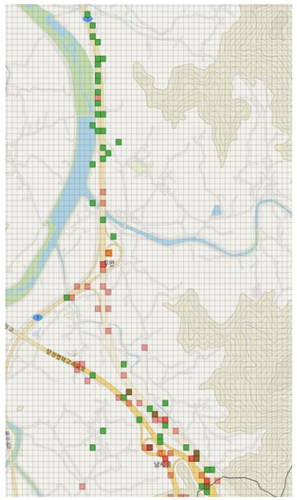
밤
낮

○ 안영IC

- 낮 :
안영IC 대전 방면의 고속도로에서 밤에 비해 낮에 사고가 다수 발생

Hotspot Grid 시각화 - 세종

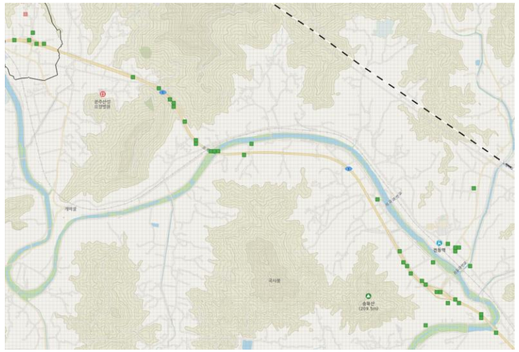
연도별



2022년
2020년, 2021년

○ 남세종IC 시청방향

- 2020년, 2022년 :
남세종IC에서 사고 다수 발생
- **2021년 :**
남세종IC뿐만 아니라 세종시청 방면 고속도로에서 사고 다수 발생

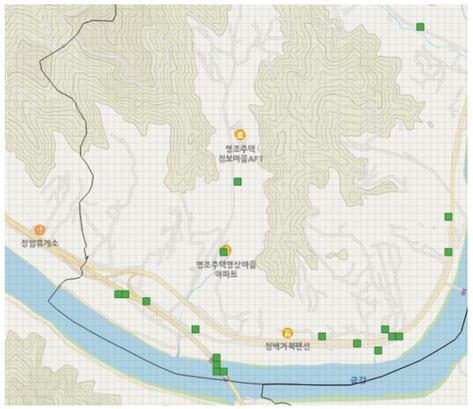


2021년
2020년, 2022년

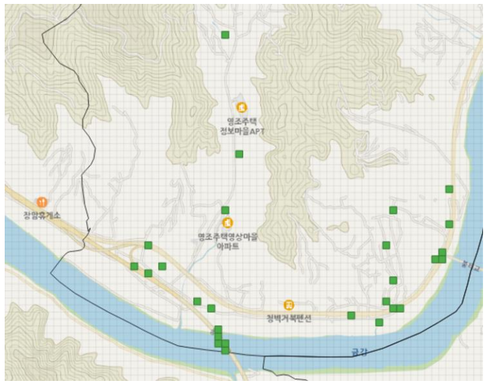
○ 홍대 세종캠퍼스

- 2020년, 2022년 :
Hotspot이 존재하지 않음
- **2021년 :**
전동면 세종로에서 사고 다수 발생

계절별



가을



낮

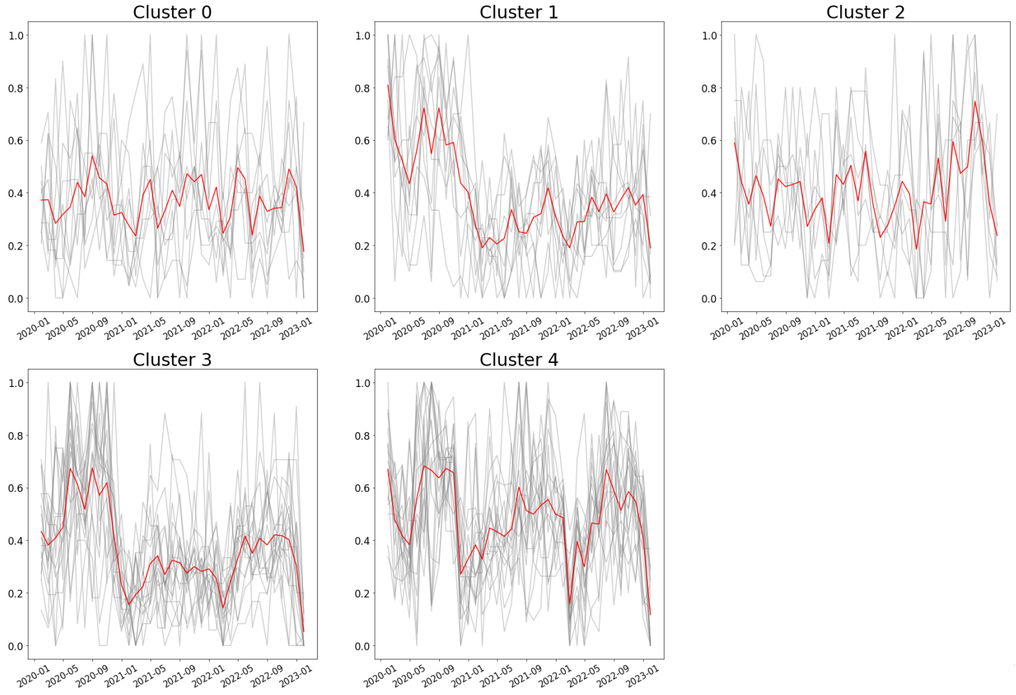
시간별

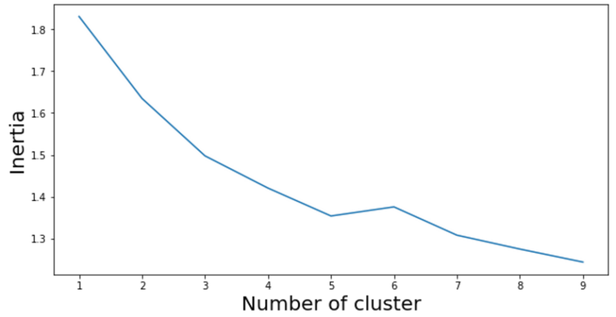
○ 장암휴게소

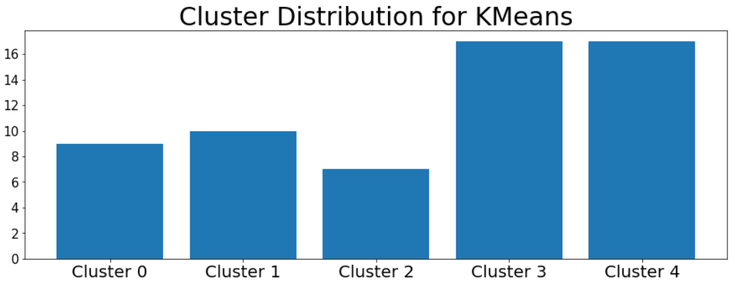
- 가을, 낮 :
세종 시청 방면 금암IC에서 사고 다수 발생

시계열 군집화 (Time Series Clustering)

시계열 군집화(Time series clustering) : 시계열 간 유사도를 바탕으로 여러 패턴의 시계열데이터를 군집화하는 방법론



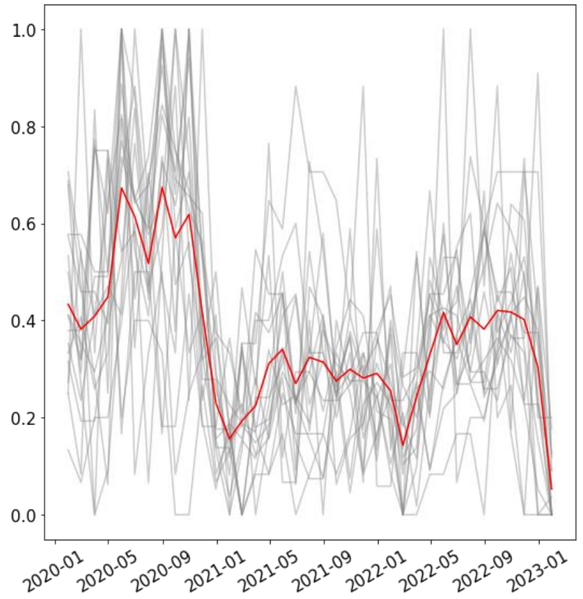




- K-means time series clustering을 통해 충남,세종,대전 별 사건발생 횟수 상위 20개의 Hotspot에 대한 월별 시계열 패턴을 군집화하여 분석
- 시계열 간 유사도를 측정하기 위하여 Dynamic Time Warping(DTW) metric을 활용
- DTW은 속도가 다를 수 있는 두 시계열 간의 유사도를 측정하는 지표
- 각 시계열을 Minmax scaler를 통해 정규화 한 후 군집화
- 군집의 개수(k)는 elbow rule을 활용하여 5개로 선정

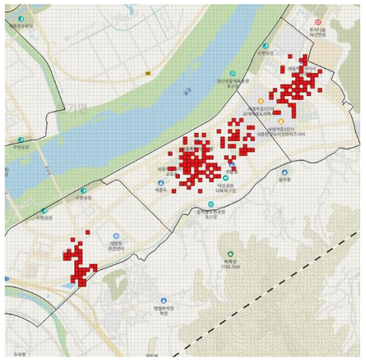
시계열 군집화 (Time Series Clustering)

Cluster 3

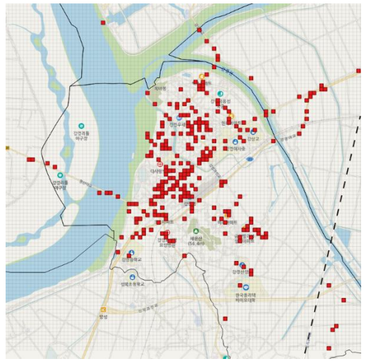


Cluster3에 포함된 Hotspot의 위치 (16개) :

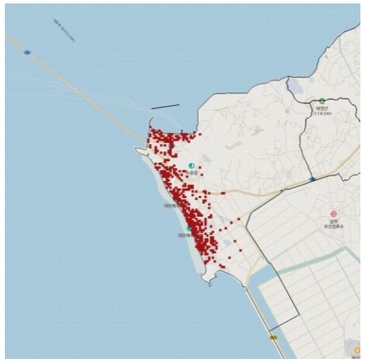
충청남도 금산군 추부면, 충청남도 청양군 청양읍
충청남도 논산시 강경읍, 충청남도 공주시 신관동
충청남도 서산시 해미면 15번 국도, 충청남도 태안군 태압읍
충청남도 보령시 신흑동 대천 해수욕장, 충청남도 아산시 권곡동
세종특별자치시 부강면, 세종특별자치시 전동면
세종특별자치시 조치원읍,
세종특별자치시 보람동 세종시청(3개)
대전광역시 유성구 덕명동
대전광역시 유성구 봉명동 유성온천역



세종 수변공원



충남 논산시 강경읍



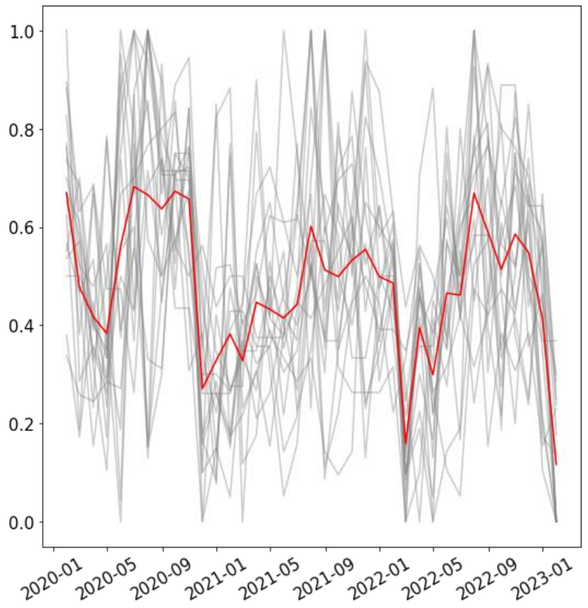
충남 보령시 대천해수욕장

○ 분석내용

- 사계절 중 여름에 사고가 가장 많이 발생
- 2020년에 사고가 가장 많이 발생
- 1년 주기로 가을에서 겨울로 변화할 때 사고가 감소하고, 겨울에서 봄으로 변화할 때 사고가 늘어나는 추세를 가짐
- 충남 논산시 강경읍에서의 봄철 유채꽃 축제는 봄철 교통사고 증가의 원인으로 추정
- 충남 보령시 대천해수욕장은 여름철 인구가 많이 몰리는 곳으로 여름에 사고가 가장 많이 발생

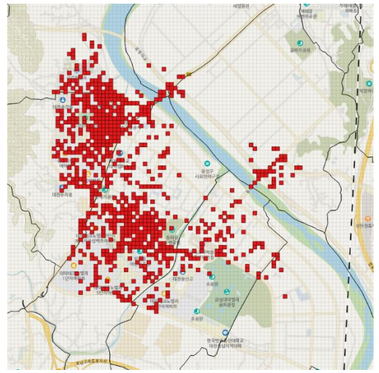
시계열 군집화 (Time Series Clustering)

Cluster 4

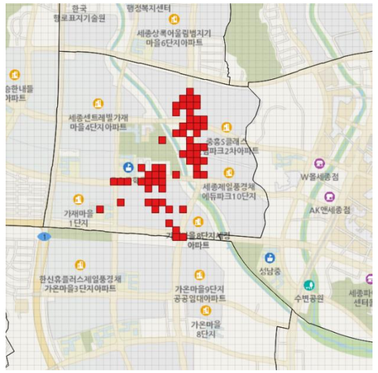


Cluster4에 포함된 Hotspot의 위치 (17개) :

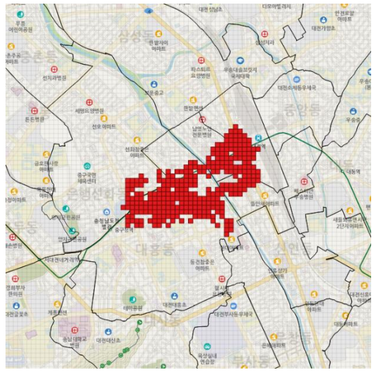
충청남도 논산시 내동, 충청남도 당진시 송산면
충청남도 서산시 석남동, 충청남도 천안시 동남구 목천읍
충청남도 서산시 동문동, 세종특별자치시 중촌동
세종특별자치시 조치원읍 , 대전광역시 동구 가오동
대전광역시 대덕구 신탄진동, 대전광역시 유성구 문지동
대전광역시 중구 중촌동, 대전광역시 유성구 신성동
대전광역시 유성구 용산동, 대전광연시 대덕구 신대동
대전광역시 중구 산성동, 대전광역시 대덕구 법동
대전광역시 중구 은행동



현대프리미엄아울렛 대전점



세종 호수공원 새뜰 근린공원



대전 스카이로드

○ 분석내용

- 2020년 5월까지 사고가 감소하고 그 후로 다시 증가한 뒤, 2020년 9월부터 다시 사고가 감소
- 2021년 겨울에서 2022년까지는 봄으로 변화할 때 사고 발생 횟수가 증가하고, 가을에서 겨울로 변화할 때 사고가 줄어드는 추세를 가짐
- 현대프리미엄아울렛 대전점의 2022년 9월 화재 사고로 인한 영업중지가 2022년 가을철 교통사고 감소의 원인으로 추정

주요 Hotspot에 대한 교통사고 발생 예측

TRAFFIC ACCIDENT PREDICTION

교통사고 발생 예측 로드맵

각 Hotspot별 교통사고 예측 모델을 생성하기 위해 필요한 변수데이터 수집 및 전처리 선제적으로 확인할 수 있는 정보만을 활용

데이터 수집, 전처리

01

AIR KOREA

기상청

Over-sampling, Feature Extraction

02

SMOTE – NC를 활용하여 over-sampling, Autoencoder를 활용하여 차원축소

RandomForest Classifier을 활용한 분류

모델링

03

성능평가

04

분류모델의 성능을 평가

데이터 수집

(기상변수)

변수명	단위
평균기온	섭씨온도(℃)
일강수량	mm
평균풍속	m/s
평균상대습도	%
평균이슬점온도	섭씨온도(℃)

(대기오염 변수)

변수명	단위
미세먼지(PM10)	$\mu g/m^3$
초미세먼지(PM2.5)	$\mu g/m^3$
오존 (O3)	ppm

- 기상, 대기환경 변수들을 활용하여 각 Hotspot들에 대한 일별 교통사고 발생 여부 예측 모델을 생성하기 위한 데이터를 수집
- 미래의 교통사고 발생 여부를 예측할 수 있는 유의미한 모델을 생성하기 위하여 예보를 통해 선제적으로 확인할 수 있는 정보만을 변수로 활용
- 기상 변수들은 기상청 기상자료개방포털[2], 대기환경 변수들은 에어코리아[3]를 통해 수집
- 기상자료개방포털은 종관기상관측장비를 통해 수집된 각 시군에 대한 일자별 기상 데이터를 제공
- 에어코리아는 각 동별로 설치된 측정소별 대기환경 데이터를 제공
- 2020년 1월 1일 ~ 2023년 1월 31일 기간에 대한 데이터 수집

데이터 전처리

- 이전 교통사고 Hotspot 분석에서 활용한 데이터셋을 Groupby 연산을 통해 일자별 교통사고 발생 횟수에 대한 데이터셋으로 변환

접수완료일시	위도	경도	군집
2020-01-01 16:08:12	126.52498	36.47439	3
…	…	…	…

노이즈 제거,
군집별로 데이터셋 분할

일자	교통사고 발생 횟수
2020-01-01	1
…	…

Cluster1

일자	교통사고 발생 횟수
2020-01-01	0
…	…

Cluster2

…

- 각 군집별 주소 및 일자를 기준으로 기상데이터, 대기환경데이터, 일자별 교통사고 발생 횟수 데이터셋 병합
- 기상, 대기환경 변수에 대한 결측치는 스플라인 보간법(Spline interpolation)을 통해 보간
- 대기환경 정보는 미세먼지 예보 등급에 따라 예보가 이루어지므로 이에 따라 대기환경 변수 범주화

(미세먼지 예보 등급)

미세먼지 농도 (㎍/㎥, 일평균)	좋음	보통	나쁨	매우 나쁨
PM10	0 ~ 30	31 ~ 81	81 ~ 150	151 이상
PM2.5	0 ~ 15	16 ~ 35	36 ~ 75	76 이상

(오존(O3) 예보 등급)

예보구간	좋음	보통	나쁨	매우나쁨
예보농도(ppm)	0~0.030	0.031~0.090	0.091~0.150	0.151 이상

데이터 전처리

- 일자를 활용하여 월(month), 일(day), 계절, 요일, workday(공휴일 여부) 변수 생성
- 교통사고 발생 횟수를 통해 종속변수인 교통사고 발생 여부 변수 생성
- 최종적으로 아래의 설명변수들을 통해 주요 Hotspot 10개에 대한 일자별 교통사고 발생 여부를 예측하는 모델을 학습
- 전체 데이터셋 중 80%를 훈련 데이터셋, 20%를 테스트 데이터셋으로 활용, 종속변수의 비율을 고려하여 sampling

설명변수

- 평균기온
- 일강수량
- 평균 상대습도
- 평균 이슬점온도
- 월(month)
- 일(day)
- PM10 (0:좋음, 1:보통, 2:나쁨, 3:매우나쁨)
- PM2.5 (0:좋음, 1:보통, 2:나쁨, 3:매우나쁨)
- O3 (0:좋음, 1:보통, 2:나쁨, 3:매우나쁨)
- 계절
- 요일
- Workday

종속변수

교통사고 발생 여부
(0 : 교통사고 발생 X / 1 : 교통사고 발생)

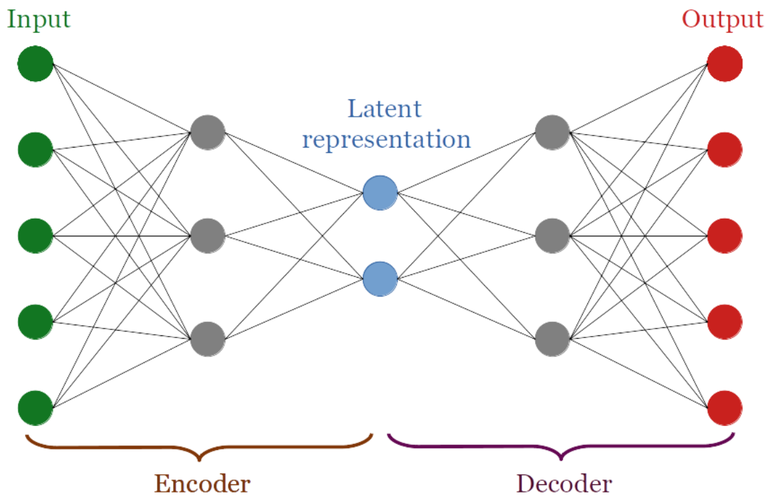
Over-sampling, Feature Extraction

Over-sampling

- 분류 예측 모델을 생성하는 10개의 Hotspot 중 종속변수의 비율에 불균형이 존재하는 경우가 다수 존재
- 데이터 불균형 문제를 해결하기 위해 SMOTE-NC(Synthetic Minority Over-sampling Technique for Nominal and Continuous) 알고리즘을 활용하여 카테고리형 변수와 수치형 변수가 혼합된 훈련 데이터셋을 Over-sampling
- Over-sampling 된 훈련 데이터셋과 테스트 데이터셋에서 명목형 변수인 월(month), 일(day), 계절, 요일, workday 변수를 One-hot Encoding

Autoencoder

- Autoencoder는 입력층과 출력층이 동일한 구조를 가지는 비지도학습의 인공신경망으로써 인코더와 디코더로 구성되어 있음
- 본 프로젝트에서는 Autoencoder를 사용하여 고차원의 데이터셋을 저차원의 잠재변수(Latent Variable)로 축약
- 훈련 데이터의 복원오차(Reconstruction error)를 최소화하는 목적함수를 통해 원본 데이터에 대해 높은 표현력을 가지는 저차원의 데이터셋 획득
- 테스트 데이터셋은 훈련 데이터셋을 통해 학습된 Autoencoder의 인코더를 통해 저차원의 잠재변수로 매핑(mapping)
- Minmax scaler를 통해 수치형 변수를 정규화 한 후 모델의 학습 및 매핑이 이루어짐



Autoencoder

Over-sampling, Feature Extraction

(Autoencoder hyperparameter setting)

Hyperparameter	chungnam_3	chungnam_148	chungnam_233	chungnam_345	sejong_43
Learning rate	0.0005	0.001	0.001	0.00005	0.0005
Epoch	200	500	300	100	200
Size of bottleneck layer	3	4	4	4	3
Number of hidden layer	2	3	3	3	4
Size of hidden layer	(32, 16)	(64, 32, 16)	(64, 32, 16)	(64, 32, 16)	(64, 32, 32, 16)
Hyperparameter	**daejeon_9**	**daejeon_11**	**daejeon_17**	**daejeon_215**	**daejeon_273**
Learning rate	0.001	0.001	0.001	0.001	0.0002
Epoch	300	200	300	100	100
Size of bottleneck layer	4	4	4	6	4
Number of hidden layer	3	2	2	2	4
Size of hidden layer	(32, 32, 16)	(32, 16)	(32, 16)	(32, 16)	(64, 32, 32, 16)

(Hotspot 별 위치)

chungnam_3	chungnam_148	chungnam_233	chungnam_345	sejong_43
충청남도 군산군 군산읍 금산시외고속버스터미널 인근	충청남도 보령시 대천5동 대천 해수욕장 인근	충청남도 보령시 대천1동 구대천역 문화관광단지 인근	충청남도 천안시 동남구 목천읍 천안농업기술센터 인근	세종특별자치시 새롬동 세종국가자격증 시험장 인근
daejeon_9	**daejeon_11**	**daejeon_17**	**daejeon_215**	**daejeon_273**
대전광역시 유성구 신성동 말바위어린이공원 인근	대전광역시 유성구 관평동 현대프리미엄아울렛 대전점 인근	대전광역시 유성구 전민동 엑스포코아 인근	대전광역시 중구 은행선화동 대전 스카이로드 인근	대전광역시 서구 괴정동 롯데백화점 대전점 인근

(데이터셋 별 사고발생 여부 비율 (0/1))

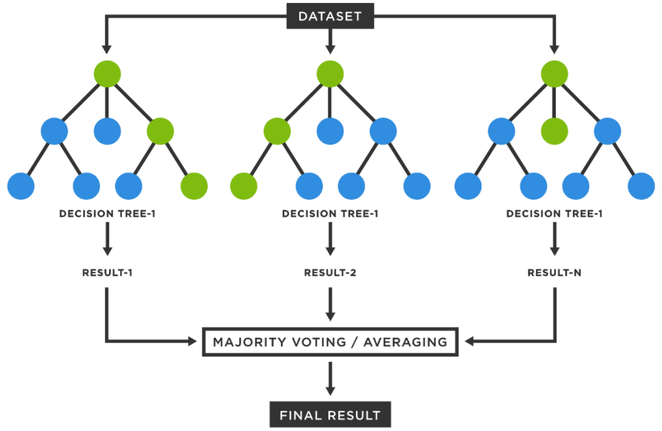
chungnam_3	chungnam_148	chungnam_233	chungnam_345	sejong_43
0.30 / 0.70	0.52 / 0.48	0.43 / 0.57	0.72 / 0.28	0.72 / 0.28
daejeon_9	**daejeon_11**	**daejeon_17**	**daejeon_215**	**daejeon_273**
0.68 / 0.32	0.21 / 0.79	0.60 / 0.40	0.21 / 0.79	0.61 / 0.39

- Autoencoder의 모든 Hidden layer에서 ReLU 활성화함수를 사용
- Adam optimizer를 사용, batch size는 100으로 설정

모델링

RandomForest Classifier

- RandomForest : 다수의 의사결정나무로부터 생성된 결과를 바탕으로 최종 예측값을 도출하는 앙상블(Ensemble) 모델
- 장점 : 과적합 문제 최소화, 대용량 데이터 처리에 효과적
- 10개의 주요 Hotspot에서 일자별 교통사고 발생 여부에 대해 예측할 수 있는 분류 모델 학습
- Hotspot별로 하나의 RandomForest Classifier를 학습



RandomForest Classifier

Hyperparameter	chungnam_3	chungnam_148	chungnam_233	chungnam_345	sejong_43
Number of trees	300	400	300	50	300
Split criterion	Gini	Entropy	Gini	Entropy	Gini
Minimum samples required to split	2	10	2	15	8
Hyperparameter	**daejeon_9**	**daejeon_11**	**daejeon_17**	**daejeon_215**	**daejeon_273**
Number of trees	50	400	500	200	150
Split criterion	Gini	Gini	Entropy	Gini	Gini
Minimum samples required to split	5	2	2	2	5

(RandomForest hyperparameter setting)

성능평가

(학습된 모델을 통해 산출된 테스트 데이터셋의 분류 예측 성능)

	0	1
Accuracy	0.61	
Precision	0.38	0.76
Recall	0.52	0.64
F1 Score	0.44	0.70

chungnam_3

	0	1
Accuracy	0.64	
Precision	0.64	0.63
Recall	0.67	0.61
F1 Score	0.66	0.62

chungnam_148

	0	1
Accuracy	0.61	
Precision	0.56	0.64
Recall	0.47	0.72
F1 Score	0.51	0.68

chungnam_233

	0	1
Accuracy	0.62	
Precision	0.81	0.38
Recall	0.62	0.61
F1 Score	0.70	0.47

chungnam_345

	0	1
Accuracy	0.61	
Precision	0.78	0.36
Recall	0.63	0.54
F1 Score	0.70	0.43

sejong_43

	0	1
Accuracy	0.63	
Precision	0.70	0.40
Recall	0.79	0.32
F1 Score	0.74	0.35

daejeon_9

	0	1
Accuracy	0.69	
Precision	0.30	0.82
Recall	0.35	0.78
F1 Score	0.32	0.80

daejeon_11

	0	1
Accuracy	0.63	
Precision	0.67	0.54
Recall	0.77	0.42
F1 Score	0.71	0.47

daejeon_17

	0	1
Accuracy	0.69	
Precision	0.30	0.82
Recall	0.35	0.78
F1 Score	0.32	0.80

daejeon_215

	0	1
Accuracy	0.61	
Precision	0.78	0.36
Recall	0.59	0.62
F1 Score	0.70	0.43

daejeon_273

- 정밀도(Precision) : 모델이 분류한 True값 중 실제 True값의 비율
- 재현율(Recall) : 실제 True값 중 모델이 True로 분류한 값의 비율
- F1 Score : 정밀도와 재현율의 조화평균

EXPECTED EFFECTS

기대효과

효율적인 교통사고 모니터링 및 관리

- HDBSCAN을 활용하여 교통사고 발생 밀도가 높은 Hotspot 지점들을 선정
- 주요 Hotspot들을 교통사고 모니터링 지점으로 선정하여 지속적인 교통사고 발생 관리 및 교통사고 방지 대책 마련에 기여

장·단기적 계획 수립

- Hotspot의 월별 시계열 패턴을 군집화하고 분석
- 시간적 요소에 따른 교통사고 발생 패턴 분석을 통해 향후 교통사고를 줄이거나 방지하기 위한 장·단기적인 정책 마련 및 전략적 계획 수립에 기여

선제적 교통사고 대응

- 기상 및 대기 예보 정보들을 활용하여 Hotspot별 교통사고 발생 예측 모델 생성
- 일별 주요 Hotspot 지점에서의 교통사고 발생을 예측하여 해당 지점에 대한 선제적 조치 및 대응 방안 마련에 기여

활용 데이터 및 참고 문헌

활용 데이터

1) Kakaomap API
(https://apis.map.kakao.com/)
2) 기상청 – 종관기상관측
(https://data.kma.go.kr/data/grnd/selectAsosList.do?pgmNo=34)
3) 에어코리아 – 최종확정 측정 자료
(https://www.airkorea.or.kr/web/last_amb_hour_data?pMENU_NO=123)

참고 문헌

참고문헌 1) Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II 17* (pp. 160-172). Springer Berlin Heidelberg.)

제1회 2023년 지역 치안 안전 데이터 분석 공모전

THANK YOU

강양이 방범대