대전·세종·충남
지역혁신플랫폼
Daejeon·Sejong·Chungnam Regional Innovation Platform

데이터 분석 최종결과보고서

목차

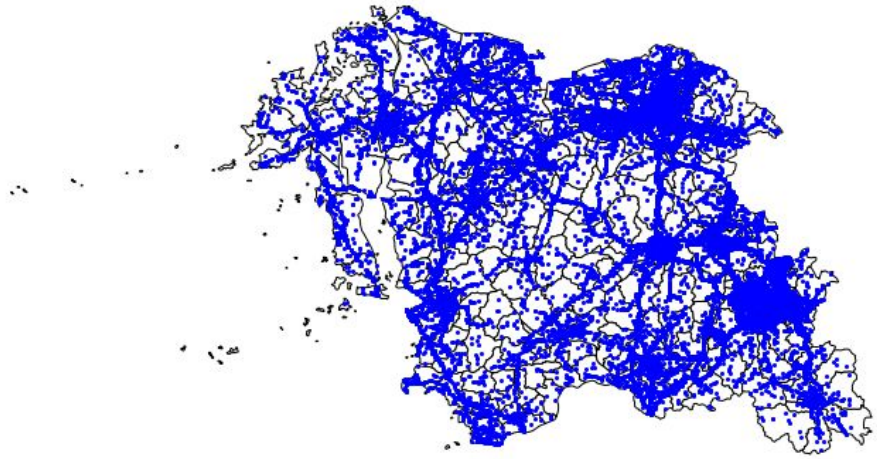
Ⅰ. 참가자 정보

제 목	대전시의 교통사고 패턴 분석 및 마코프 프로세스를 통한 사고 빈도 예측 모델 개발	
팀 명	백야컴과	
성 명	노태윤, 이재빈, 전성후, 이종훈	
연락처	휴대폰	*010-9031-9869*
	E-mail	*nrbsld@korea.ac.kr*

Ⅱ. 개요

1. 연구 범위

팀 백야컴과는 “충남·세종·대전 지역 교통사고 분석 및 예측”이라는 공모전 주제 아래 분석 및 예측을 진행하였다. 주어진 데이터 중 KP2020, 2021을 활용하여 충남, 세종, 대전에서 2021-2023년 발생한 교통사고를 시각화하였고, 결과는 다음과 같다.

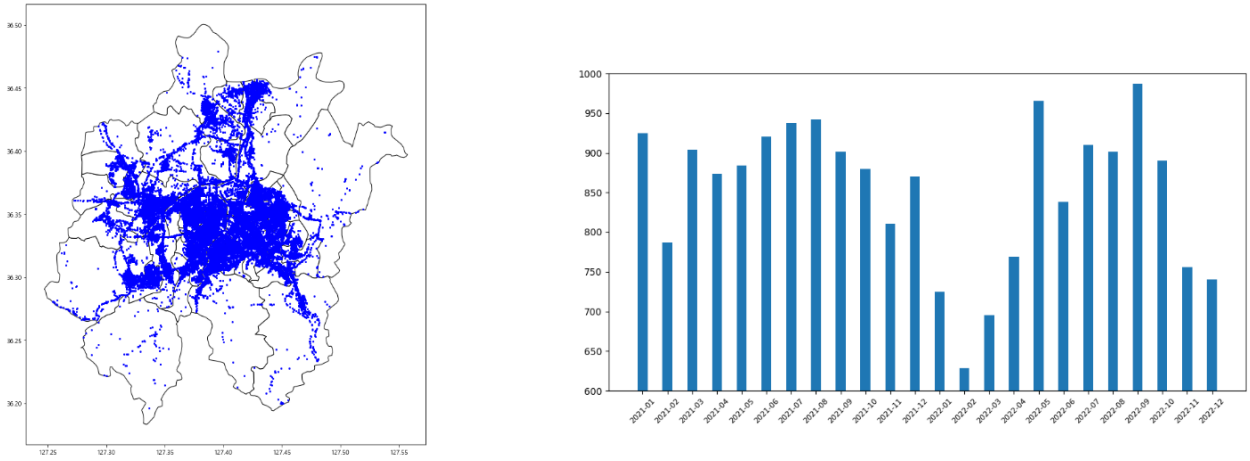


〈그림 1〉 충남, 세종, 대전 교통사고 시각화

해당 자료로 확인할 수 있듯이, 충남, 세종, 대전, 지역을 통합적으로 분석하기에는 각 지역마다의 경향성과 분포가 지나치게 다양했다. 교통사고 발생의 일반적인 경향성을 분석하기 위해, 한 지역을 선정하여 그 지역에서 발생한 교통사고를 다각적으로 분석해 보려고 한다. 분석 대상 지역으로는 행정동의 구분이 체계적이고, 데이터가 풍부한 대전특별시를 선정하였다.

데이터 분석을 위해 데이터 전처리 과정을 거쳤다. 우선 속성의 명칭을 직관적으로 바꿔주었다. 이후 시간 데이터를 초 단위로 바꿔주었으며, 모든 속성이 주어지지 않은 데이터 값은 삭제 하였다. 이후 주어진 위도, 경도 값을 바탕으로 대전에 속하는 데이터만을 필터링하였다. 필터링 된 데이터를 해당 시간대에 해당하는 대전시의 날씨 정보와 함께 “daejeon_weather.csv”로 저장하였다.

2. 연구 배경 및 목적



<그림 2> 대전시 교통사고 시각화

대전의 교통사고 발생 횟수는 2021년 11667건, 2022년 9805건으로 1년 사이에 16% 감소하였다. 또한 자동차 등록대수가 2021년 692,702대에서 2022년 707,928대로 2.18% 상승했음에도, 자동차 1만대 당 교통사고는 2021년 168명에서 2022년 139명으로 약 18% 감소하였다.(대전광역시 차량등록 사업소, 2022). 사고가 감소했다는 점은 고무적이지만, 전국 광역시·도 기준으로는 여전히 최하위권으로 관련 연구 및 예측 방안이 시급한 상황이다. (교통사고분석시스템(TAAS), 2021).

교통사고 분석에 관한 연구는 다수 존재한다. 기상요건을 중심으로 한 연구로는, 최새로나 외(2012)는 기상 및 교통조건이 고속도로 교통사고 심각도에 미치는 영향을 분석하였다. 이경준 외(2015)는 도로위의 기상요인이 교통사고에 미치는 영향을 시간당 강수량, 강수유무, 기온, 풍속을 중심으로 분석하였다. 해당 연구들은 교통사고의 발생과 기상요인이 서로 상관관계를 가진다는 것을 입증하였고, 교통사고의 예방과 예측에 있어 기상요인의 고려가 필수적이라는 사실을 보인다.

교통사고가 발생한 곳의 시공간적, 사회적 요인이 분석의 대상이 되기도 한다. Antonio(2019)는 DBS Clustering을 활용해 런던의 사고다발구역을 군집화하고, 실시간 교통사고 예측에 활용하였다. Kumar과 Toshniwal (2015)은 K-modes 군집화 기법을 사용해 인도 데라둔의 교통사고를 분석하였다. 김운용 외(2020)은 대구지역의 교통사고 패턴에 공간군집이 존재하는지와 기상요인의 영향을 파악하였다. 기상요인과 사회적 요인의 영향을 받는 사고 발생의 시공간적 위치에 대해 분석을 진행한다면, 교통사고 발생에 대한 일반적인 경향성을 찾아낼 수 있을 것으로 기대된다.

교통사고 건수 예측 모델에 관한 연구 또한 다수 진행되었다. 전통적으로 포아송 회귀모형이 사용되었으며, 최근에는 딥러닝 등을 사용한 머신러닝 기반의 예측 모델이 주목받고 있다.

포아송 회귀모형의 경우 종속변수가 평균과 표준편차를 따른다고 가정하기에 실제 상황에 적용하기 어렵다는 단점이 있다. 딥러닝 기반 예측 모델의 경우 전국이 아닌 광역시 차원의 교통사고를 분석하기 때문에 학습 데이터의 양이 충분치 않고, 분석 과정에서 그 인과 과정을 알기 어렵다는 단점이 있다.

이를 감안하여 교통사고의 빈도수를 예측하기 위한 모델로 마코프 프로세스를 선택하였다. 교통사고의 발생을 예측하기 위한 모델 중, 교통사고 빈도수를 현실과 유사하게 예측하고 그 인과관계를 외연적으로 분석할 수 있는 최적의 방안이라고 판단하였다. 마코프 프로세스를 이용하여 예측 모델을 개발하고, 실제 발생 빈도수와 비교를 하여 정확도 및 성능을 검증할 예정이다. 대전시의 교통사고 빈도를 정확하게 예측을 할 수 있다면, 이 방법론을 세종, 충남 등 다른 지역에도 확대 적용하여, 도로별, 시기별로 그 위험도와 사고 발생 빈도를 예측하여 체계적인 관리를 할 수 있을 것으로 기대된다.

Ⅲ. 분석/시각화 결과 상세내용

1. 기상요인에 따른 교통사고 빈도 분석

(1) 분석 방식의 목적 및 이론적 배경

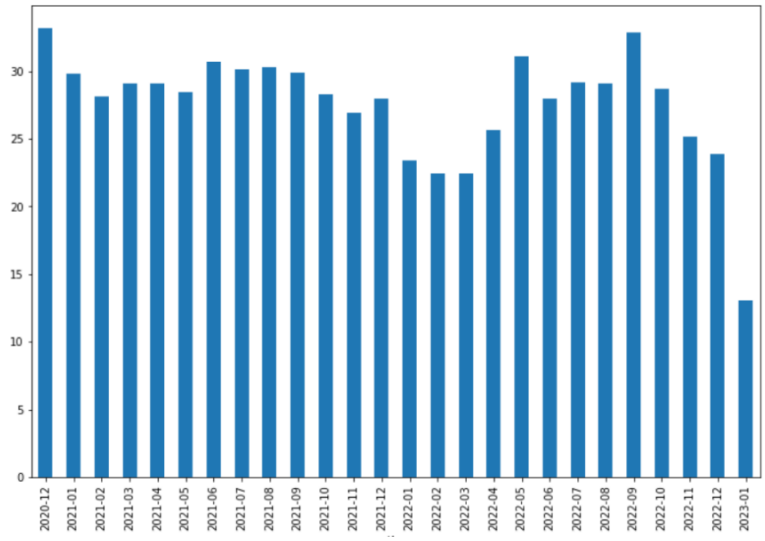
교통사고의 분석에 있어 기상요인은 가장 많이 고려되는 요인 중 하나이다. 일반적으로는 비나 눈이 오는 이상기후 상태에서 사고가 많이 발생하고, 강수량과 적설량이 많아질수록 교통사고의 증가로 이어질 수 있다고 생각된다. 그러나 Yannis, Karlaftis(2010)의 연구에서는 더 나쁜 기상조건이 운전자들이 더욱 조심스럽게 운전하도록 유도하여 오히려 교통사고의 발생을 감소시킨다는 safety-offset hypothesis(안전 상쇄 가설)를 제안하여 기상 상태에 따른 운전자의 심리 상태도 교통사고 발생에 영향을 미칠 수 있다고 설명하였다. 김현욱 외(2013)는 우리나라에서의 기상과 교통사고의 연관성을 분석하여 우리나라에서도 강수와 적설이 교통사고의 빈도와 음의 상관관계를 가지며, 해당 가설이 적용될 수 있음을 시사했다. 본지에서는 대전의 교통사고 발생 양상을 기온, 습도, 풍속, 강수 유무, 강수량 등의 다양한 측면에서, 운전자의 심리적 요인과 함께 분석하여 해당 요인들이 교통사고 발생 빈도에 미치는 영향을 살펴본다. 분석에는 집계 기간에 해당하는 '기상청 날씨누리'의 대전 지역 '지역별상세관측자료'를 30분 단위로 가져와 사용하였다.

(2) 분석 내용 및 결과

가. 월별 일평균 교통사고 발생 분석

월별로는 2022년 1, 2, 3, 11, 12월의 일간 평균 발생횟수가 현저히 적게 나타났다. 일원분산분석을 실시하여 이것이 통계적으로 유의미한 차이인지를 검정해보고자 하였다. 분석 결과, 평균 차이는 유의미하였다($F=5.934$, $p<0.05$). Bonferroni correction을 이용하여 사후검정을 실시한 결과, 2022년 1, 2, 3, 12월에서 유의미한 평균 차이가 있었다.

	표준편차	평균
2020년 12월	7.27	33.19
2021년 1월	6.42	29.81
2021년 2월	4.09	28.11
2021년 3월	6.38	29.06
2021년 4월	6.97	29.10
2021년 5월	6.71	28.45
2021년 6월	6.24	30.67
2021년 7월	7.56	30.16
2021년 8월	5.92	30.32
2021년 9월	5.83	29.93
2021년 10월	7.07	28.29
2021년 11월	5.65	26.90
2021년 12월	6.79	28.00
2022년 1월	6.97	23.39
2022년 2월	6.14	22.46
2022년 3월	5.29	22.42
2022년 4월	5.90	26.52
2022년 5월	6.14	31.06
2022년 6월	7.42	27.93
2022년 7월	7.13	29.16
2022년 8월	6.11	29.06
2022년 9월	6.76	32.90
2022년 10월	7.95	28.71
2022년 11월	7.60	25.13
2022년 12월	4.98	23.87



	df	sum_sq	mean_sq	F	PR(>F)
C(month)	24.0	6085.837722	253.576572	5.963712	5.030615e-17
Residual	735.0	31252.141226	42.519920	NaN	NaN

표 2, 월별 평균 일간 교통사고 발생 횟수의 일원분산분석 결과

	12월	01월	02월	03월	04월	05월	06월	07월	08월	09월	10월	11월	12월	01월	02월	03월	04월	05월	06월	07월	08월	09월	10월	11월	12월
2020년 12월		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
2021년 1월	FALSE		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
2021년 2월	FALSE	FALSE		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2021년 3월	FALSE	FALSE	FALSE		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2021년 4월	FALSE	FALSE	FALSE	FALSE		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2021년 5월	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2021년 6월	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
2021년 7월	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2021년 8월	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
2021년 9월	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
2021년 10월	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2021년 11월	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2021년 12월	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2022년 1월	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE		FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
2022년 2월	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE		FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
2022년 3월	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE
2022년 4월	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2022년 5월	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
2022년 6월	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2022년 7월	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE		FALSE	FALSE	FALSE	FALSE	FALSE
2022년 8월	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE		FALSE	FALSE	FALSE	FALSE
2022년 9월	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	TRUE	TRUE
2022년 10월	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	FALSE
2022년 11월	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE		FALSE
2022년 12월	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	

표 3, Bonferroni correction을 이용한 사후 검정 결과 (TRUE = (p < 0.05)

교통사고분석시스템(TAAS)의 2015-2021년 월별 교통사고 사고건수 데이터에 따르면 12 - 3월, 즉 겨울철에 전국적으로 교통사고가 덜 발생하였다. 분석 기간 동안 대전광역시의 월별 교통사고 발생 횟수에서 나타난 평균 차이도 이런 경향성에 따라 나타난 것으로 보인다. 송남기(1990)의 연구에 의하면, 이는 겨울철은 운전자들이 기상이변으로 인해 교통 환경이 좋지 못하는 사실을 염두에 두고 안전운전을 하고, 추운 날씨로 인해 차량통행량이 다른 계절에 비해 적기 때문으로 분석된다.

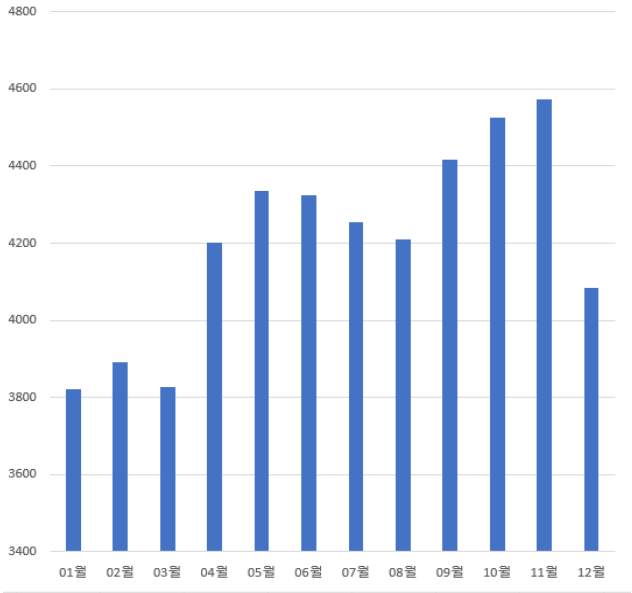


그림 10 2015 - 2021년 전국 월별 교통사고 사고건수

나. 기온에 따른 교통사고 발생 빈도

기온을 섭씨 5도 크기의 구간으로 나누어, 해당 구간에서 30분 간 발생할 것으로 예상되는 교통사고의 수를 계산하였다. 계산에 이용된 식은 다음과 같다.

$$N_A = \frac{E_A}{W_A}$$

E_A는 집계기간동안 구간 A에서 발생한 사고의 수를 의미하며, W_A는 집계기간동안 구간 A에 해당했던 단위시간의 수를 의미한다. 단위시간은 30분이며, 구간은 섭씨 -15도부터 35도까지 5도 단위로 설정하였다.

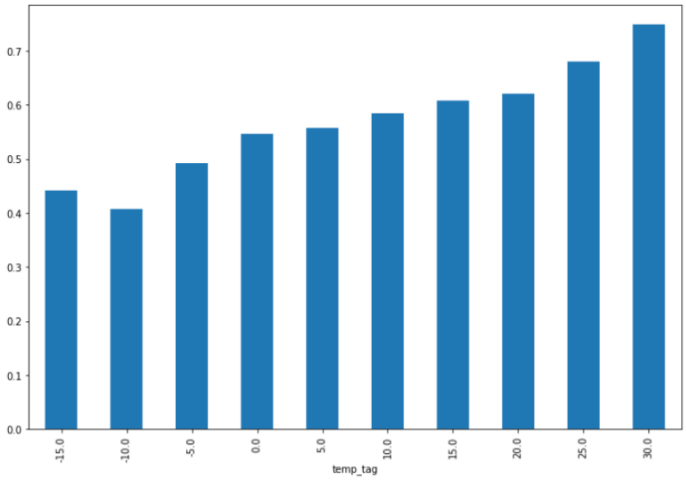


그림 11. 기온에 따른 교통사고 발생 빈도

기온이 높을수록 사고의 발생 빈도가 증가하는 것으로 나타났다. 현대해상화재보험 교통기후환경

연구소 (2021)의 연구 결과에 따르면, 기온이 섭씨 30도 초과일 때 타이어펑크사고가 66% 증가하고, 기온과 직접적인 연관이 있는 불쾌지수가 80초과인 경우 사고가 15% 증가하는 양상을 보였다. 해당 분석에 사용된 데이터로 사고 유형을 판단할 수 없지만, 기온 상승에 따른 '스탠딩 웨이브' 현상에 따른 타이어펑크 증가와, 높은 불쾌지수로 인한 운전자의 심리적 요인이 영향을 미치는 것으로 보인다. 또한 (1)에 언급되었듯 겨울철의 교통사고 발생 빈도가 상대적으로 낮게 나타나기 때문에, 기온이 낮을 때 교통사고 발생 빈도가 더 낮게 측정되는 것으로 보인다.

다. 습도에 따른 교통사고 발생 빈도

상대 습도를 섭씨 5도 크기의 구간으로 나누어, 해당 구간에서 30분 간 발생할 것으로 예상되는 교통사고의 수를 계산하였다.

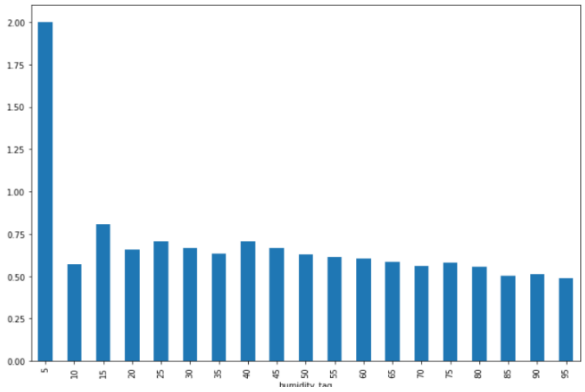


그림 12 습도에 따른 교통사고 발생 빈도

습도 0-5%에 해당하는 구간은 해당 습도 구간에 해당했던 30분간 두 번의 사고가 발생한 것으로 이상치로 판단하여 분석에서 제외한다. 전반적으로 습도에 따른 교통사고의 발생 빈도는 큰 차이를 보이지 않았고, 습도가 증가할수록 교통사고가 감소하는 양상은 보였으나 그 정도는 크지 않았다.

라. 풍속에 따른 교통사고 발생 빈도

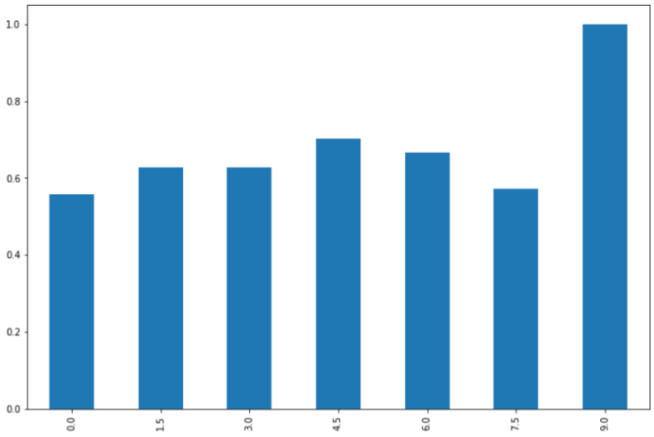


그림 13 풍속에 따른 교통사고 발생 빈도

풍속을 1.5m/s 크기의 구간으로 나누어, 해당 구간에서 30분 간 발생할 것으로 예상되는 교통사고의 수를 계산하였다.

풍속 9m/s 이상에 해당하는 구간은 해당 풍속 구간에 해당했던 시간이 2시간 30분뿐이었기 때문에 이상치로 의심되고, 해당 구간을 제외하면 풍속에 따른 교통사고의 발생 빈도는 큰 차이를 보이지 않았다. 그러나 김영덕 외(2007)의 연구 결과에 따르면 정지 상태, 차종에 따라 차이를 보이지만 약 13m/s 이상의 풍속에서는 차량의 영향에 크게 영향 받을 수 있다고 알려져, 매우 강한 풍속에서의 교통사고 발생 빈도와 관련된 부분에 대한 연구를 위해서는 더 많은 표본이 필요할 것으로 보인다.

마. 강우 여부에 따른 일간 교통사고 발생 차이

강우 여부에 따른 일간 교통사고 발생 횟수의 평균 및 표준편차는 다음과 같다.

	평균	표준편차
강우 有	28.41	5.75
강우 無	27.52	7.58

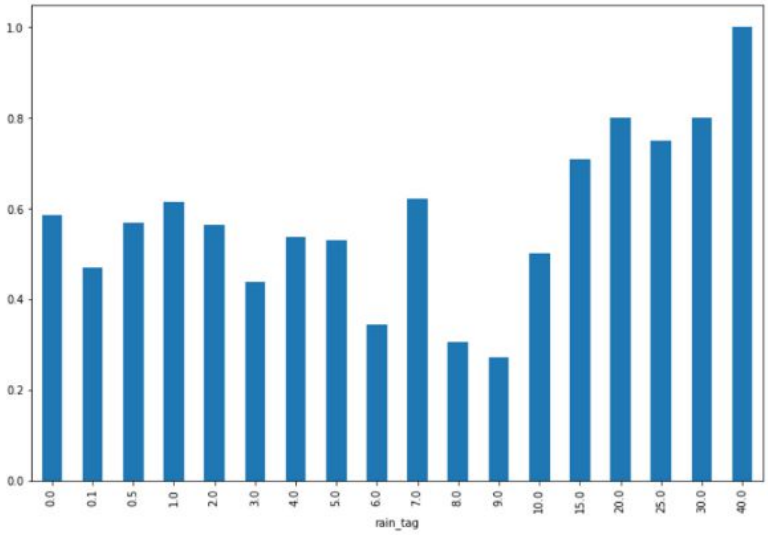
표 4 강우 여부에 따른 일간 교통사고 발생 횟수 통계

두 집단에 T-검정을 진행한 결과, 등분산 가정시 p=0.17, 비등분산 가정시 p=0.11로, 두 집단의 평균 간에 유의미한 차이가 존재하지 않음을 확인할 수 있었다. 즉, 강우 유무는 교통사고 발생 횟수에 큰 영향을 미치지 않았다.

	df	sum_sq	mean_sq	F	PR(>F)
C(month)	1.0	97.088074	97.088074	1.844333	0.17484
Residual	774.0	40744.350070	52.641279	NaN	NaN

<그림 14> 강우 여부에 따른 두 집단의 T-검정 결과

바. 시간 당 강수량에 따른 교통사고 발생 빈도



〈그림 15〉 강수량에 따른 교통사고 발생 빈도

강수량을 기상청 레이더에서 제공하는 강수량 구간으로 나누어 해당 구간에서 30분 간 발생할 것으로 예상되는 교통사고의 수를 계산하였다.

10mm/h까지는 큰 차이를 보이지 않았으나, 기상청 기준 '강한 비'에 해당하는 15mm/h 이상의 강수량일 때에는 그 이하일 때 보다 사고 빈도가 30분당 0.2회 이상 높은 경향을 보였다. 그러나 그 양상이 비례적으로 나타나지는 않았으며, 오히려 강수량이 6-9mm/h일 때가 비가 오지 않을 때보다 사고가 더 적게 나타남을 확인할 수 있었다. 이는 운전자들이 나쁜 기상조건에서 더 조심히 운전하여 오히려 교통사고가 더 적게 일어난다는 Yannis, Karlaftis(2010)의 safety-offset 가설을 보충한다. 또한, 15mm/h 이상의 폭우에서는 운전자의 심리적 요인보다 노면의 미끄러움, 시야방해 등 물리적 요건이 사고 발생에 더 큰 요인으로 작용하여 사고가 증가한다고 추측할 수 있다.

2. 교통사고 군집별 교통사고 분석

(1) 분석 방식의 목적

본 연구에서는 대전 전체 교통사고를 일괄적으로 분석하였다. 그러나 교통사고의 발생은 도로의 통행량, 지역의 사회경제적 여건, 지형별 특징 등 구역별로 발생 양상과 원인이 상이하다. 이를 보완하고자, 군집화를 활용하여 교통사고가 자주 발생하는 구역을 추출하고자 한다. 이후 발생 양상과 날씨 등 외부 변인들과의 연관성을 분석하여 특정 구역에서의 사고 원인을 더욱 정밀하게 연구하고자 한다.

(2) 이론적 배경 및 분석방법 선정 근거

군집화는 DBSCAN Clustering 모델을 사용하여 진행하였다. DBSCAN Clustering은 대규모의 데이터에 적용하기 적합한 밀도 기반의 군집화 알고리즘이다. 여기서 “밀도 기반”은 데이터를 군집화할 때 데이터가 밀집해 있는 곳을 찾겠다는 것을 의미한다. 군집화는 데이터간의 거리가 충분히 가까우면, 그 데이터를 군집에 포함시키는 것으로 이루어진다. 예를 들어, 세지점 A,B,C에서 사고가 발생했다고 하자. A와 B간의 거리가 충분히 가깝다면, A와 B는 하나의 군집에 속하게 된다. 또한 B와 C의 거리도 충분히 가깝다면, A와 C의 거리가 멀더라도 B와 C는 한 군집으로 묶이게 되어 A, B, C는 같은 군집에 속하게 된다. 이처럼 DBSCAN Clustering은 단순히 가까운 데이터들을 묶는 것에 그치지 않는데. 서로 연결되어 있는 데이터 집단을 발견하기에 유리하고, 군집으로부터 멀리 떨어진 데이터를 배제할 수 있어 분석의 정확도가 높다. 교통사고는 통행량이 많은 교차로 근방이나 혼잡한 상권 주위에서 발생하는빈도수가 높기에 본 연구에서는 DBSCAN Clustering이 교통사고 군집을 추출하는 데 적합하다고 판단하여 활용하였다.

(3) 연구 방법

가. 군집 산정 과정

KP2020,2021의 위도-경도 자료를 활용하여 2020.12.01~ 2023.01.17 기간의 사고 지점을 시각화하였다. 또한 통계청에서 제공하는 행정동 경계 파일을 통해 대전 내의 도로들도 시각화하였다.

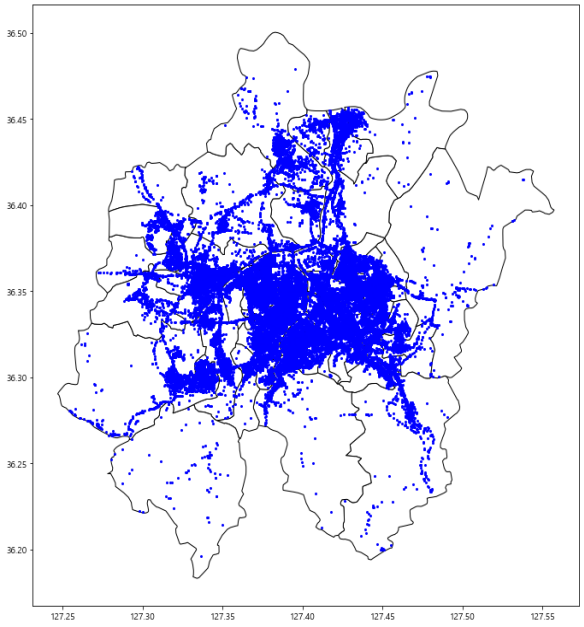


그림 16). 대전 교통사고 지점 시각화

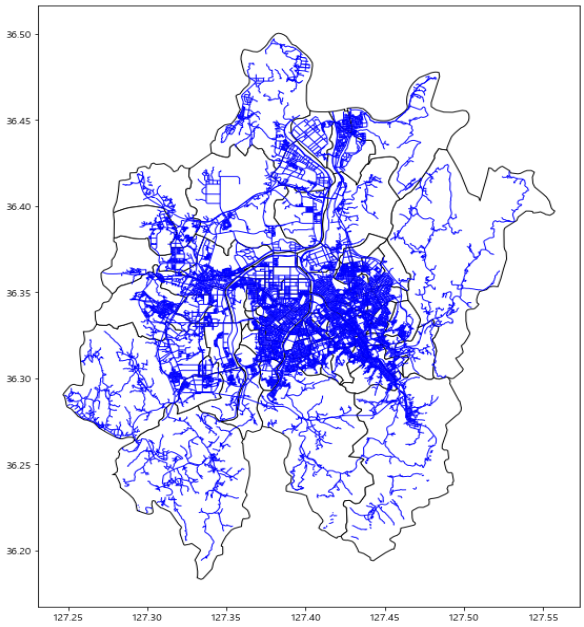
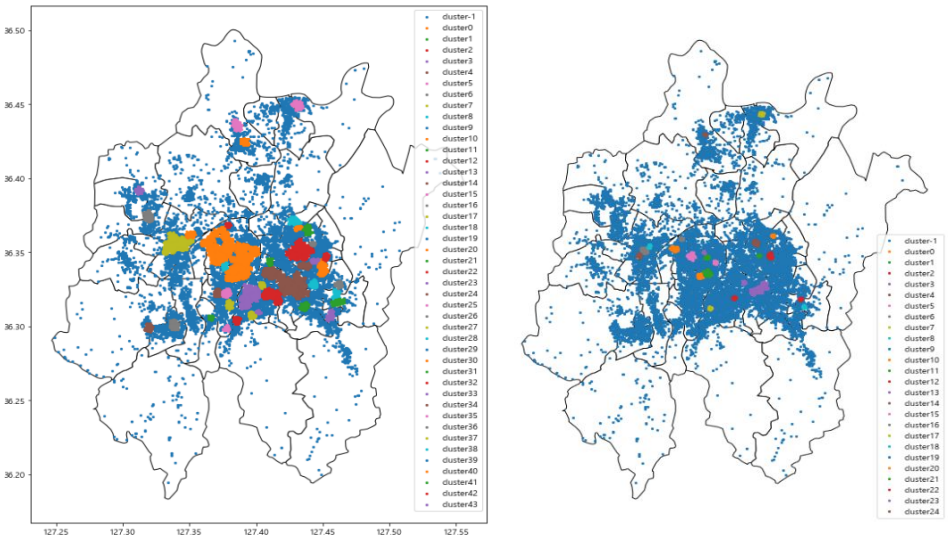


그림 17). 대전 도로 시각화

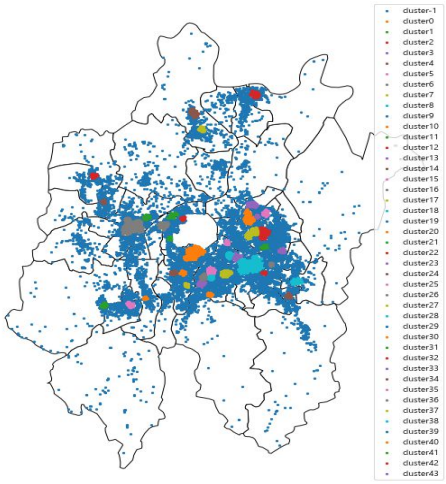
자료와 같이 도로 밀도가 높은 둔산동, 대전역 부근의 번화가는 교외 지역에 비해 교통사고의 발생건수가 높고, 사고 장소가 밀집되어 있다. DBSCAN Clustering을 진행하기 위해서는 군집 속 데이터 간의 최대 거리인 eps를 지정해 주어야 한다. 특징이 다른 교외, 시내에 동일한 eps값을 적용하여 군집화를 진행할 경우 시내에 규모가 비정상적으로 큰 군집이 형성되거나 교외에 군집이 잘 형성되지 않을 수 있다.

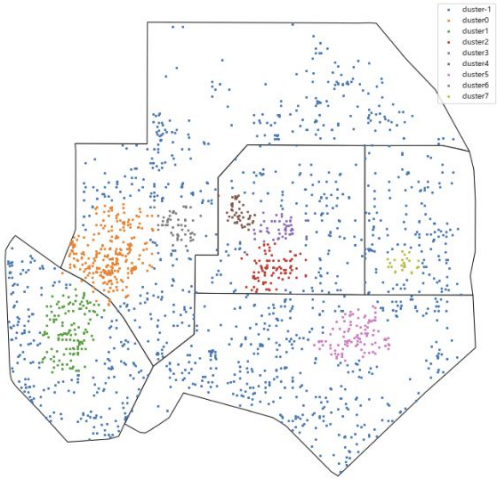


〈그림 18, 19〉 eps=0.05 eps=0.015일때의 군집화 결과

실제로 대전시에 일괄적으로 eps값을 지정해 군집화를 진행할 경우, 그림 3과 같이 둔산동 근방에 규모가 큰 군집이 형성되고, eps값을 줄일 경우 교외 지역은 군집화가 거의 이루어지지 않는 문제가 발생하였다.

문제점을 극복하기 위해 본 연구에서는 대전의 대표적 번화가인 둔산동, 갈마동, 탄방동과 대전 여타 구역(교외)을 분리하여 군집화를 진행하였다. 두 그룹의 eps값과 군집으로 묶이기 위한 데이터의 최소 개수인 min_samples를 도로, 사고발생구역의 밀집된 정도에 따라 다르게 설정하였다. 그 결과는 다음과 같다; 둔산, 탄방, 갈마동 eps는 0.15, min_samples=25로 설정하였고, 교외 지역은 eps=0.04, min_samples=40으로 지정하였다.





〈그림 20〉 둔산, 탄방, 갈마 제외 군집화 결과　　〈그림 21〉 둔산, 탄방, 갈마동 군집화 결과

결과적으로 둔산, 탄방, 갈마 3동과 대전 여타 지역에서 각각 7개, 43개의 군집을 추출하였다. 군집에 포함된 사고의 건수는 다음 표와 같다;

이름	표본 수
-1	1212
0	295
1	134
2	93
3	48
4	49
5	103
6	57
7	25

이름	표본 수	이름	표본 수	이름	표본 수
-1	12927	14	66	29	54
0	754	15	150	30	61
1	117	16	238	31	73
2	327	17	165	32	50
3	158	18	93	33	73
4	153	19	77	34	45
5	107	20	60	35	50
6	761	21	66	36	43
7	246	22	63	37	52
8	1141	23	145	38	57
9	116	24	64	39	43
10	307	25	83	40	41
11	82	26	93	41	41
12	160	27	77	42	42
13	115	28	157	43	40

<표 5> 군집별 표본수. -1은 어느 군집에도 속하지 않음을 의미한다.

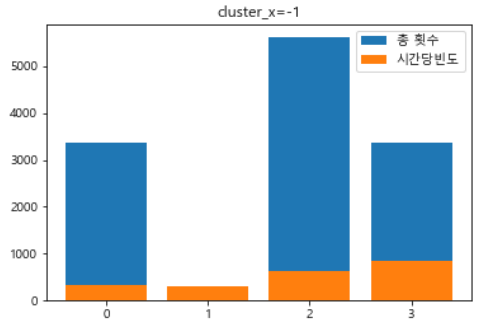
나. 군집별 분석결과

군집별 분석의 목적은 사고가 발생 빈도가 높은 구역을 군집화를 통해 찾아내고 특징을 밝히는 것이다. 교통사고의 발생 양상을 파악하여 예방하는데 활용할 수 있고, 경향성 파악을 통한 예측도 가능하다. 이를 위해 본 연구에서는 시간대별 교통사고 건수의 분석을 통해 군집별 교통사고의 양상을 확인하였다. 또, 선형회귀 기법을 통해 분기별 교통사고 수를 시간에 따라 군집별로 분석하였다. 군집의 수가 많기에, 보고서에 모두 나타내기 어려움이 있어, 타 군집과 동일한 경향성을 보이거나 통계적으로 무의미한 그래프는 나타내지 않았다.

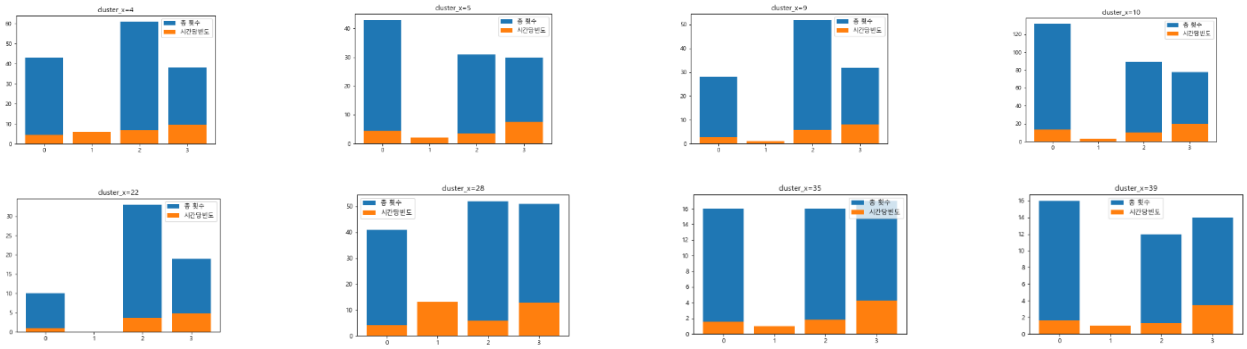
● **시간대별 분석**

교통사고 건수를 전날 23:00~금일 7:59의 심야시간대, 8:00~8:59까지의 출근시간, 9:00~17:59까지의 낮 시간대, 18:00~22:59까지의 퇴근 및 귀가시간대로 분류해 계수하였다. 각 시간대는 그래프 x축에 순서대로 0, 1, 2, 3으로 표기하였다. 그리고 시간대별 교통사고의 건수와 그 건수를 시간대의 시간 길이로 나누어 시간당 빈도도 구해 막대그래프로 나타내었다.

[교외 결과]



<그림 22>. 군집에 포함되지 않은 모집단의 결과 그래프



<그림 23~30>. 군집별 시간대별 교통사고 총 횟수/시간당 빈도. 왼쪽 위부터 cluster 4, 5, 9, 10, 왼쪽 아래부터 cluster 22, 28, 35, 39이다.

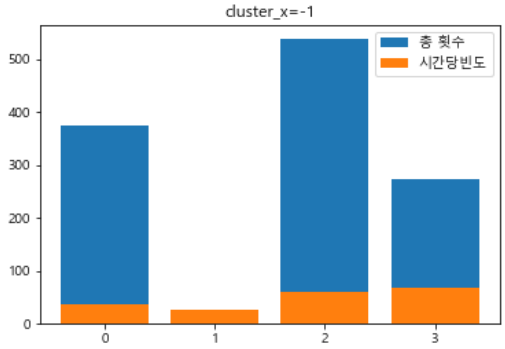
교외 지역의 경우, 군집에 속하지 않은 모집단의 경우 출근 시간의 시간당 사고 빈도가 가장 적었고, 퇴근 및 귀가 시간대에 사고 빈도가 가장 높았다. 이 경향성은 대부분의 군집에서 유지되었다. 예외적으로 cluster 28의 경우, 출근시간의 사고 빈도 13건으로, 퇴근, 귀가시간의 빈도인 12.77에 비해 높았다. 원인으로는 Cluster 28이 위치한 판암역 인근에 출근시간에 교통량이 많은 옥천로가 존재하고, 대전 판암초등학교 및 아파트 단지가 다수 밀집해 있어 출근 시간에 교통이 매우 혼잡하기 때문인 것으로 추측된다.

그림 31). Cluster 28의 위치

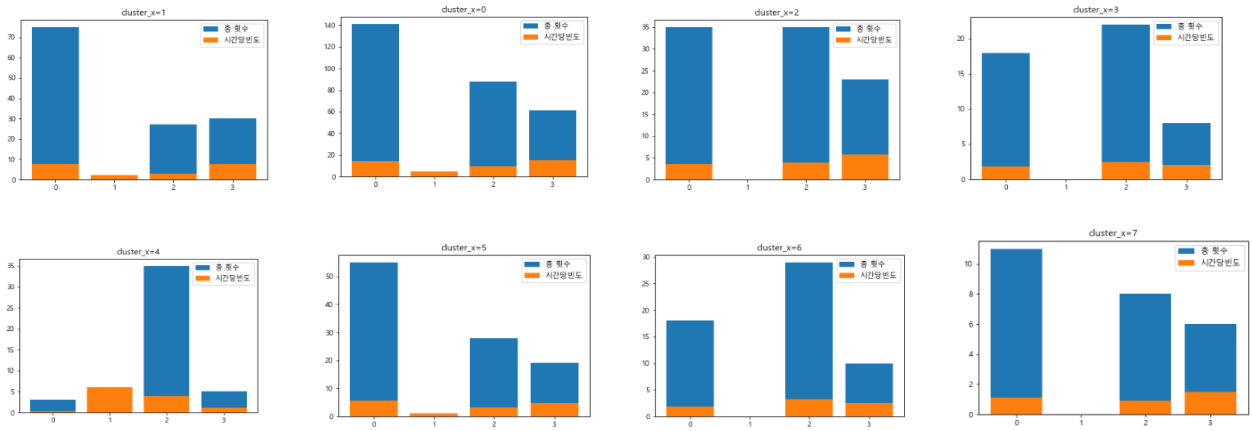
그림 32). 판암역 인근 지도

실제로 그림 32를 보면, 옥천로 주위에 아파트 단지가 5개나 밀집해 있음을 알 수 있다. 이와 관련해서는 판암역 인근 시간대별 교통량 분석 등을 통해 위 가설을 검증할 필요가 있어 보인다.

[둔산, 탄방, 갈마동 결과]



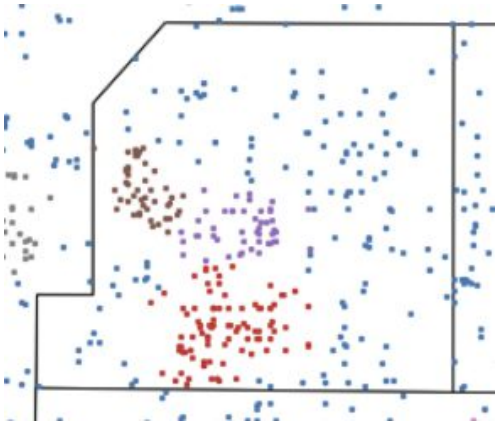
<그림 33> 군집에 포함되지 않은 모집단의 결과 그래프



<그림 34~41>. 군집별 시간대별 교통사고 총 횟수/시간당 빈도. 왼쪽 위부터 cluster 0, 1, 2, 3, 왼쪽 아래부터 cluster 4, 5, 6, 7이다.

둔산, 탄방, 갈마동의 군집에 포함되지 않는 모집단은 교외 지역의 그림 22와 거의 유사한 경향을 보였다. 반면, 둔산, 탄방, 갈마동의 군집들은 시간당 빈도 수가 출근시간 제외 모든 시간대에서 비슷하거나, 군집들 중 절반이 출근시간대의 교통사고가 0건일 정도로 교외 구역과 차이를 보였다. 특이한 점은 대전지방검찰청-시청역 사이의 cluster 4, 5, 6중 cluster 4만 출근시간대의 시간당 사고 빈도가 매우 높고, 낮 시간대의 사고가 대다수라는 점에서 주변의 다른 군집들과 차이

를 보였다.



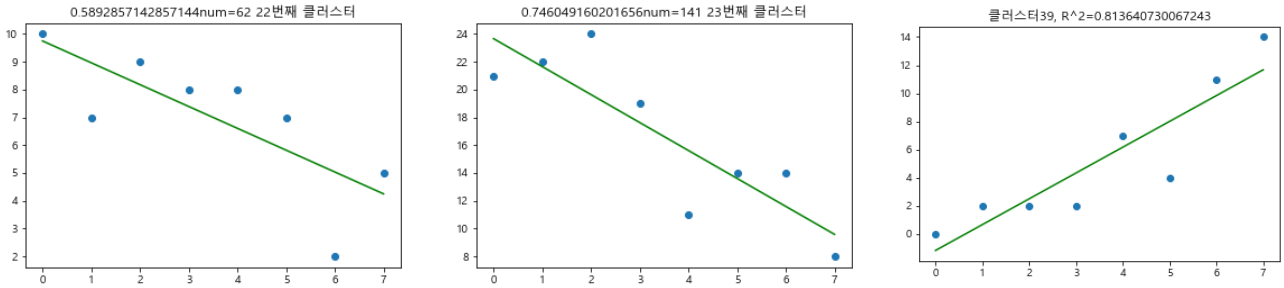
<그림 42> cluster 4, 5, 6의 모습.
갈색이 4, 빨강이 5, 보라색이 6이다.

<그림 43> 검찰청-시청역 사이 지도

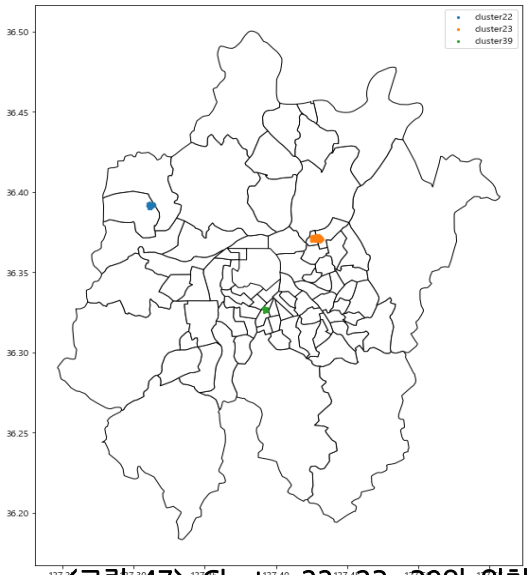
동일한 지리적, 교통 환경에 위치한 cluster 5, 6과 확연히 구별되는 특징이라는 점에서 위 특징들은 cluster 4만의 고유 특성으로 생각되어진다. 추후 cluster 4, 5, 6에 속하는 사고들에 대한 면밀한 원인 분석을 통해 이와 같은 고유 특성을 검증하여 사고 예방에 활용할 필요가 있어 보인다.

● 분기별 분석

2021.01.01.~2022.12.31.까지의 기간을 3개월씩 총 8분기로 나누어 각 분기별 군집별로 속하는 교통사고의 수를 분석하였다. 분기는 순서대로 x축에 0부터 7까지의 수로 나타내었고, y축에는 각 분기별 교통사고의 수를 나타내었다. 그리고 선형회귀 기법을 적용하여 각 군집별 사고의 증감 경향을 확인할 수 있는 통계적으로 유의미한 군집들을 추출하였다. 통계적 유의미함은 상관계수 값으로 판단하였다. 위 방법론을 적용하여 교외 구역에서 3개의 증가 또는 감소 경향을 보이는 군집을 추출하였다. 둔산, 탄방, 갈마동 군집에선 통계적으로 유의미한 경향을 발견하지 못했다.



<그림 44~46>. Cluster 22, 23, 39의 분기별 사고 수 그래프.



<그림 47> Cluster 22, 23, 39의 위치
Cluster 22-파랑, 23-오랜지, 39-초록

<그림 44~46>과 같이 Cluster 22, 23은 분기별 교통사고 건수가 감소 추세에, Cluster 39는 증가 추세에 있다. <그림 3>과 같이, 대전시의 교통사고는 분기에 따라 주기적으로 증가와 감소를 반복한다. 그러나 이런 주기적인 증감에도 불구하고 특정 구역의 분기별 교통사고가 지속적으로 증가 또는 감소하면, 그 구역은 대전시 여타 지역에 비해 교통사고가 강하게 증가 또는 감소하는 특징을 지녔다고 볼 수 있을 것이다. 이와 같이 선형회귀를 사용하여 여러 사고다발구역의 사고의 증감을 확인한다면, 증감에 비례하여 경찰력을 재배치하는 등의 조정을 통해 앞으로 발생할 교통사고에 효율적으로 대비할 수 있을 것으로 기대된다.

Ⅳ. 교통사고 빈도 예측 방안 모델링

1. 이론적 배경

마코프 체인(마코프 프로세스)이란 한 상태(state)에서 다른 상태로 전이할 때 특정한 확률적인 특성을 따르는 것을 의미한다. 이 중요한 특성은 현재 상태에서 다음 상태로 넘어갈 때 현재 시점보다 이전의 과거 상태에는 의존하지 않는다는 것이다. 즉, 상태 간 전이가 오로지 이전 n개의 상태에 의존하여 이루어지는 것을 말한다. 예를 들면, 초기 숫자 3이 주어지고, 주사위를 던져서 2 이하의 눈이 나오면 초기 숫자에 1을 빼고 3 이상 4 이하의 눈이 나오면 그대로 두고 5 이상 6 이하의 눈이 나오면 1을 더하는 시행을 반복한다고 해보자, 여러 번의 시행 끝에 초기 숫자가 6이 되었다. 다음 시행으로 초기 숫자가 7이 될 확률은 1/3이고 이는 현재 이전 시행의 기록에 영향을 받지 않는다. 이를 마코프 체인을 따르는 사건이라고 한다.

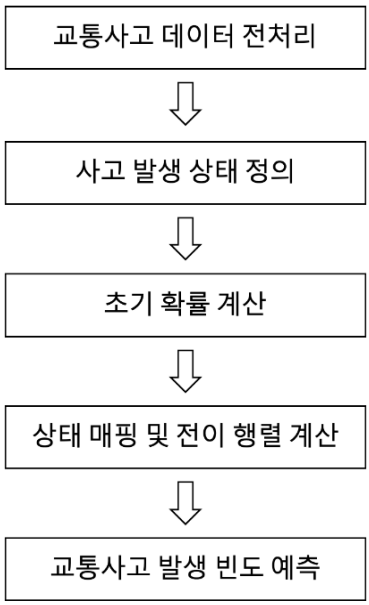
마코프 체인을 수식적으로 살펴보면 다음과 같다. 시간에 따라 변화하는 확률 변수는 $X(t)$ 로 하고, 이 값이 나타내는 시간이 임의의 시점 $t_0 < t_1 < \ldots < t_n < t_{n+1}$ 일 경우에 $X(t)$ 가 이산값이면 다음과 같다.

$$P = P(a < X(t_{n+1}) = x_{n+1} | X(t_n) = x_n, \ldots, X(t_1) = x_1)$$
$$= P(X(t_{n+1}) = x_{n+1} | X(t_n) = x_n)$$

여기서 $t_1, \ldots, t_{n-1}$은 과거 시점이고 t_n은 현재, t_{n+1}은 미래를 나타낸다. 과거(현재 이전) 상태들이 어떠한 상태를 가지더라도 확률값은 오로지 현재와 미래 상태에만 의존한다는 것을 나타낸다. 이때 이 시간 값이 이산값이면 마코프 체인이라고 한다.

마코프 체인은 가능한 상태들을 집합으로 생성한 '상태 집합', 정의된 상태들이 초기 상태에 가질 수 있는 발생 확률인 '초기 확률', 각 상태 간의 전이확률로 구성된 '전이 행렬'로 구성되어 있다. 본 연구에서는 이 3가지 구성요소를 정의함으로써 대전 지역의 교통사고 발생 빈도 예측 모델을 구축하고 적용하였다.

2. 예측 모델 구축 및 검증



먼저 대전 지역의 교통사고 발생 빈도를 예측하기 위한 2년간(2021~2022) 월별 교통사고 발생 빈도 데이터를 수집하여 코딩하였다. 2022년 10월, 11월 12월의 교통사고 빈도를 예측하기 위해서 2021년 1월부터 예측하는 달의 직전 달까지의 데이터를 필터링하였다. 그리고 다음과 같은 과정을 거쳐서 교통사고 발생 빈도를 예측하였다.

첫 번째 단계는 사고 발생상태 정의이다. 이 단계에서는 교통사고 발생 빈도 상태 집합을 구분하는 것이 필요한데, 교통사고 발생 빈도에 따른 위험 정도를 상태(S)라 하며, 상태집합은 교통사고 발생 빈도의 위험 상태가 될 수 있는 값들의 범위를 나타낸다. 본 연구에서는 교통사고 발생 빈도에 따른 위험 정도에 따라 3단계인 주의 상태(S_1), 경계 상태(S_2), 심각 상태(S_3)로 구분하였다. 상태 집합을 구분하기 위해서 월별 교통사고 발생 빈도 데이터 분석을 하여 평균과 표준편차를 구했다.

N(조사 개월 수)	평균	표준편차	최솟값	최댓값
23	856.52	89.37	629	987

〈표 1〉 월별 교통사고 발생 빈도 데이터 분석 결과

상태 집합을 구분하기 위한 임계값은 전체 데이터를 3개의 그룹으로 구분했으므로 평균값을 구한 후, 여기에 일정한 값을 더하거나 뺀 두 값을 임계값으로 설정하고자 하였다. 따라서 수집한 데이터의 평균값에서 표준편차를 더하거나 뺀 다음, 1을 더하거나 빼서 값을 약간 보정하는 방식으로 설정하였다.

상태 집합 구분	주의 상태(S_1)	경계 상태(S_2)	심각 상태(S_3)
임계값 범위	< 767	767 ~ 945	945 <

〈표 2〉 교통사고 발생 빈도별 상태 집합 구분

두 번째 단계는 초기 확률 계산이다. 초기 확률은 정의된 각 교통사고의 발생 빈도의 위험 상태가 초기에 발생할 수 있는 확률로써, 비교적 최근에 발생한 교통사고 빈도수를 이용한다. 교통사고가 얼마만큼 발생할 것인가 예측하기 위해 최근 발생한 사고 건수가 어느 정도의 영향을 주는지 정의하는 단계이다. 본 연구에서는 최근 5개월의 데이터를 이용하여 초기 확률을 구하기로

하였다.

$$P(\frac{a}{T},\frac{b}{T},\frac{c}{T}) = P(S_1,S_2,S_3) \quad \text{식(1)}$$

$$T = \sum_{i=1}^{3} T_i = a + b + c \quad \text{식(2)}$$

식(1)의 a,b,c는 최근 교통사고 발생 빈도 데이터 중에서 앞서 정의한 각 상태 집합에 해당하는 횟수이고, T는 식(2)에서 알 수 있듯이 각 상태 집합의 횟수를 모두 더한 데이터의 총 개수이다. 이 단계에서는 T = 5이다. 따라서 초기 확률값은 최근 교통사고 발생 빈도의 각 상태 집합에서 전체 상태 집합의 합으로 나누어 주면 된다. 최근 발생한 교통사고에서 산출된 초기 확률값은 총합이 1이 되어야 하며, 즉 식(3)의 조건을 만족한다.

$$\sum_{i=1}^{3} P(S_i) = 1 \quad \text{식(3)}$$

초기 확률을 산출하기 위한 최근 5개월 (2022년 7월 ~ 2022년 11월) 교통사고 발생 빈도에 따른 상태 집합을 구분한 결과는 <표 3>과 같다.

최근 5개월	2022 7月	2022 8月	2022 9月	2022 10月	2022 11月
발생빈도	910	901	987	890	756
상태 집합	S_2	S_2	S_3	S_2	S_1

<표 3> 최근 5개월('22.7月~11)간 교통사고 빈도 상태 집합 구분

최근 5개월 상태 집합 횟수는 S_1는 1, S_2는 3, S_3는 1로 나타났다. 따라서 초기 확률은 다음과 같다.

$$P = (0.2, 0.6, 0.2)$$

세 번째 단계는 상태 매핑(mapping) 및 전이 행렬 계산이다. 위험 정도에 따른 상태 간의 전이 행렬을 구한다. 이를 산출하기 위해 세 가지 절차를 거쳐야 한다. 첫째, 각 위험 상태별 교통사고 발생빈도에 따라 상태 집합과 매핑해준다. 앞서 초기 확률을 구하기 위해서 했던 최근 5개월 데이터 교통사고 빈도 상태 집합 구분을 이제는 전체 데이터에 해주는 것이다. 둘째, 매핑된

상태에서 하나의 상태에서 다른 상태로 전이되는 횟수를 구하여 전이 행렬을 생성한다. 데이터가 n개라면, 전이 행렬의 모든 요소들의 합은 n-1이다. 셋째, 생성된 전이 행렬을 확률행렬로 바꾸어주기 위해서 식(4)와 같이 각 상태의 전이 행렬의 확률을 산출하고, 이는 식(5)의 조건을 만족해야 한다. 식(4)와 같은 확률행렬을 구하기 위해서 정의된 상태 집합의 전이 횟수를 전체 전이 횟수로 나누어 확률을 구했다.

$$P = \begin{vmatrix} P_{11} & P_{12} & P_{13} & \dots & P_{1n} \\ P_{21} & P_{22} & P_{23} & \dots & P_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ P_{n1} & P_{n2} & P_{n3} & \dots & P_{nn} \end{vmatrix} \qquad \text{식(4)}$$

$$P_{ij} \geq 0, \sum_{j=1}^{n} P_{ij} = 1 \,(\text{단}, i = 1,2,\dots,n) \qquad \text{식(5)}$$

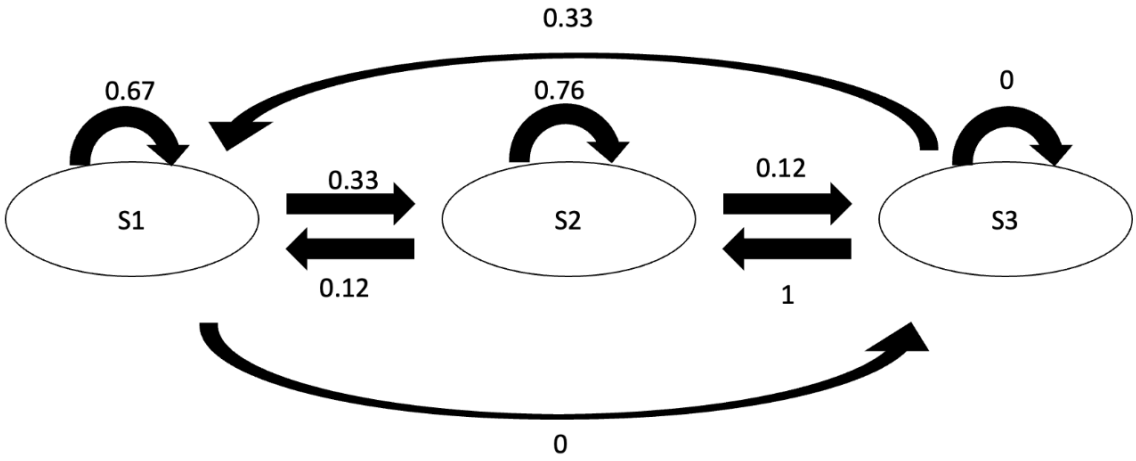
전이 행렬을 확률로 나타내면 확률행렬이 되는데 확률행렬은 확률벡터를 열로 가지는 정사각행렬이다. 확률벡터란 음수가 아닌 성분들로 이루어졌으며, 그 성분들의 합이 1인 벡터를 말한다. 정의에 의하면 우리가 구한 확률행렬의 열들의 합이 1이 되어야 하지만, 우리는 전이 행렬을 구할 때, P_{ij} 를 'i' 상태에서 'j' 상태로의 전이로 두었기 때문에 행과 열이 뒤바뀐 전치행렬이 나왔다. 그러나 확률행렬은 정사각행렬이므로 행렬곱 연산 결과 행렬의 크기도 동일하고 이후 계산 과정에서 원행렬과 전치행렬의 계산 결과가 같기 때문에 우리는 이대로 연구를 진행했다.

수집한 교통사고 발생 빈도 데이터를 상태 집합 임계값의 범위에 따라 매핑하고 전이 행렬을 생성하면 다음과 같다. 오른쪽은 전이 행렬의 횟수를 확률로 계산하여 전이 행렬로 나타낸 것이다.

	S_1	S_2	S_3
S_1	2	1	0
S_2	2	13	2
S_3	0	2	0

	S_1	S_2	S_3
S_1	0.67	0.33	0
S_2	0.12	0.76	0.12
S_3	0	1	0

〈그림 1〉 전이 행렬(2022.12)



〈그림 2〉 교통사고 발생 빈도 상태전이 다이어그램(2022.12)

마지막 단계는 교통사고 발생 빈도 예측이다. 빈도 예측은 전 단계에서 산출된 초기 확률과 교통사고 발생상태 전이 행렬을 이용하여 교통사고가 발생할 빈도 확률과 앞으로 발생할 교통사고 빈도를 예측한다. 식(6)은 식(1)과 식(4)를 이용한 교통사고 발생 확률식이다.

$$P(S_j) = \sum_{i=1}^{n} P(S_i) P_{ij} \qquad \text{식(6)}$$

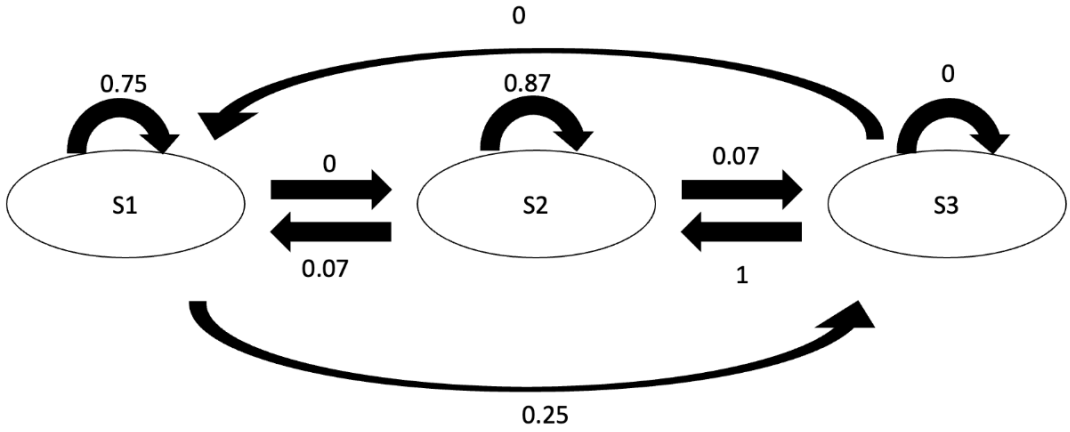
여기서 n은 교통사고 발생상태의 집합 수이고, $P(S_i)$는 교통사고가 발생할 수 있는 초기 확률이며, P_{ij}는 전이 행렬의 값이다.

이제 앞에서 구한 초기 확률과 발생 확률 전이 행렬을 이용해서 2022년 12월의 교통사고 빈도 예측을 해보자. 전체, 최근 6개월, 최근 1년간 데이터의 평균을 가지고 예측하고 빈도수의 최댓값을 가지고 예측을 해본 결과는 다음과 같다.

구분	전체 평균: 856.5	6개월 평균: 880.3	1년 평균: 827.9	최댓값: 987
예측빈도	618.41	635.60	690.48	823.16
예측률(%)	83.6	85.9	80.8	96.3

〈표 4〉 2022. 12 교통사고 빈도 예측 (실제 관측값: 740)

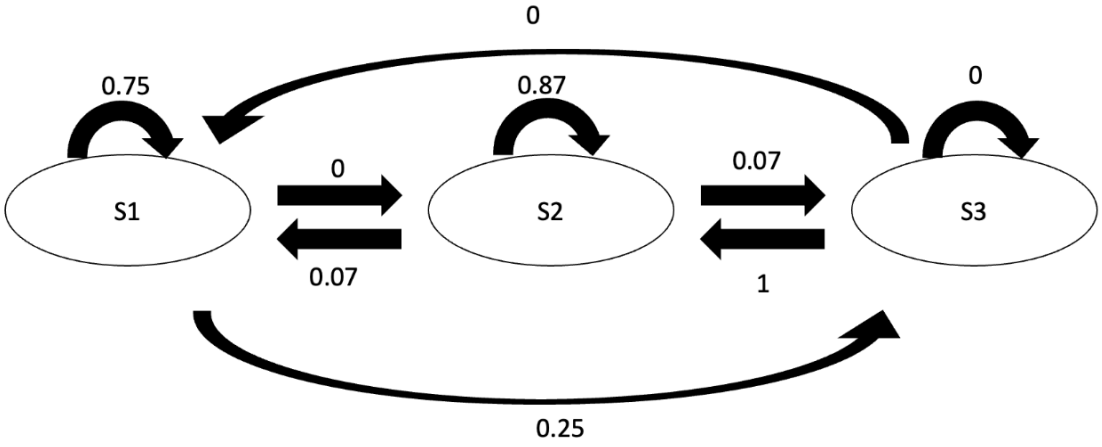
평균적으로 실제 관측값과의 오차가 15%정도 된다는 분석 결과가 나왔다. 그런데 최댓값을 기준으로 했을 때, 오차가 5% 미만 수준으로 나온 것을 인지하고 유의미한 결과인지 확인하기 위해서 똑같은 절차를 걸쳐서 2022년 11월과 2022년 10월의 교통사고 빈도를 예측했다.



〈그림 3〉 교통사고 발생 빈도 상태전이 다이어그램(2022.11)

구분	전체 평균: 861.1	6개월 평균: 915.2	1년 평균: 832.4	최댓값: 987
예측빈도	771.54	635.60	690.48	823.16
예측률(%)	102.1	108.5	98.7	117.0

〈표 5〉 2022.11 교통사고 빈도 예측 (실제 관측값: 756)



〈그림 4〉 교통사고 발생 빈도 상태전이 다이어그램(2022.10)

구분	전체 평균: 859.7	6개월 평균: 895.0	1년 평균: 831.6	최댓값: 987
예측빈도	792.67	825.19	766.72	910.01
예측률(%)	89.06	92.71	86.15	102.24

〈표 5〉 2022.10 교통사고 빈도 예측 (실제 관측값: 890)

분석 결과 2022년 11월 데이터에서 평균을 이용한 예측 수치들은 오차율 10% 미만으로 비교적 정확한 예측이 나왔지만, 최댓값을 이용한 예측은 오차율 17%로, 앞서 2022.12월 예측에서 최댓값을 이용한 예측이 정확하게 나온 것에서 유의미한 내용을 도출할 수는 없었다.

Ⅴ. 결론 및 기대효과

대전시의 교통사고 관련 수치들은 감소 중이지만 교통사고는 인명과 직접적 관련이 있고, 심각한 사회경제적 피해를 끼치기에 분석 및 예방이 필요하며, 이를 위해 과거의 데이터를 면밀히 분석 할 필요가 있다.

본 연구에서는 주어진 데이터를 바탕으로 교통사고에 기상요인이 어떤 영향을 미치는 지에 대한 분석을 실시하였다. 기상요인 중 기온이 높을수록 사고가 더 많이 발생한다는 사실이 드러났고, 이는 운전자의 불쾌지수 증가, 노면의 고열로 인한 타이어펑크 사고 증가 등으로 추정된다. 습도, 풍속은 교통사고의 발생에 큰 영향을 끼치지 않는 것으로 보이나, 표본의 부족으로 인해 추가적인 연구가 필요할 것으로 보인다. 강우여부의 경우 역시 큰 영향을 끼치지 않았고, 강수량은 오히려 비가 올 때는 사고 빈도가 감소하다, 폭우 시에 사고 빈도가 증가하는 경향성을 보였다. 이는 safety-offest 가설로 설명할 수 있다. 일반적인 강우 시에는 운전자가 더 안전하게 운전하려는 경향을 보이고, 폭우 시에는 그런 경향으로도 상쇄할 수 없는 물리적 요인으로 인해 사고 발생이 증가하는 것으로 보인다.

나아가 군집화를 통해 사고 발생 빈도가 높은 특정 구역을 발견하고, 교통사고 발생 양상을 분석하였다. 먼저 DBS Scanning을 활용해 교외에서 44개의 군집, 둔산, 탄방, 갈마동에서 8개의 군집을 추출하였다. 그리고 교통사고의 건수를 사고가 일어난 시간대별, 분기별로 나누어 분석하였다.

시간대별 분석의 경우, 교통사고 거의 모든 군집에서 퇴근 및 귀가 시간대(오후 6시~10시 사이)의 시간대별 사고 빈도가 높았고, 출근 시간에 가장 낮았다. 교외 구역과 둔산, 탄방, 갈마동 모두 경향에서 크게 어긋나는 군집들이 존재하였으며, 각 시간대별로 빈도에서 확연한 차이를 보였던 교외 군집과 달리 번화가 군집에서는 출근시간을 제외하고 거의 균등한 빈도를 보이기도 했다.

분기별 분석의 경우, 선형회귀를 통해 분기에 따라 사고의 증감이 뚜렷한 3개의 군집을 추출할 수 있었다.

두 가지 분석 방식을 통해 군집별로 어느 시간대에 사고가 자주 발생하는지 파악할 수 있었고, 어느 군집이 지속적으로 사고가 증가/감소하는지도 알 수 있었다. 분석 내용을 바탕으로 군집별로 사고가 가장 잦은 시간대에, 사고가 지속적으로 증가하는 군집에서 경찰력을 투입하는 등의 방안을 통해 효율적으로 교통사고에 대응할 수 있는 역량을 갖추리라 기대된다.

그 후 미래의 교통사고 발생 빈도를 예측하기 위해 마코프 프로세스를 2년간의 데이터에 적용하였다. 예측 모델은 다음을 통해 구축하였다.

1. 평균값에서 표준편차를 더하고 뺀 값을 이용하여 임계값을 설정하고, 이를 바탕으로 교통사고를 분류하였다.
2. 최근 대전시에서 발생한 5개월간의 교통사고를 상태로 매핑한 후 초기 확률을 산출하였으며, 전이 행렬을 구하였다.
3. 전이 행렬과 확률을 곱하여 최종 확률 값을 구하였다.

4. 구축된 모델을 통해 S2 상태의 확률이 가장 높게 나왔고, 실제로 발생하였음을 알 수 있었다
5. 예측된 빈도수와 실제 빈도수를 비교한 결과 2022년 10월, 11월, 12월 모두 95% 이상의 정확도로 예측할 수 있음이 확인되었다.
6. 마코프 체인 예측 모델을 통해 교통사고 발생 빈도를 미리 예측할 수 있으며, 사고 발생 확률에 따라 체계적인 관리 및 예방이 가능함을 알 수 있었다.

그러나 명확한 한계점도 존재한다. 사건 빈도수를 예측한 것이기에, 특정 장소 및 시간에서의 사건 발생 확률을 예측하기에 적합하지 않았다. 또한 한 달 단위로 예측을 하기에, 앞으로의 경향성을 파악하기 단점도 있다. 이외에도 임계값과 유사한 교통사고를 지속적으로 보이는 경우, 임계값의 설정에 따라 빈도수 예측값이 큰 차이를 보인다는 문제점도 있다. 이를 보완하기 위해 임계값을 설정하는 더욱 명확한 기준이 필요하며, 한계점을 보완할 필요가 있을 것이다. 더욱 정교화 되고 세밀한 구축 모델을 발전시킨 후, 적용 범위를 좁혀 특정 지역, 행정동을 바탕으로 한다면 교통사고 예측 및 사전 예방을 위한 시스템으로 사용될 수 있을 것으로 가대한다.

Ⅳ. 활용 데이터 및 참고 문헌 출처

1. David C. Lay., Steven R. Lay., Judi J. McDonald. Linear Algebra and Its Applications. Pearson, 프로텍 미디어. 2017.
2. 장은진. "마코프 체인을 활용한 해양 사고 빈도 예측에 관한 연구". 한국해양경찰학화보 10, no.3 (2020): 145 ~ 170.
3. "마르코프 체인 정의 및 예시", Data to Impact | 데이터 사이언스 미국 유학/취업 티스토리, 2020년 12월 12일 수정, 2023년 2월 15일 접속, https://cosmy.tistory.com/entry/%EB%A7%88%EB%A5%B4%EC%BD%94%ED%94%84-%EC%B2%B4%EC%9D%B8-Markov-Chain-%EC%A0%95%EC%9D%98-%EB%B0%8F-%EC%98%88%EC%8B%9C.
4. 송남기. (1990). 레미콘 트럭의 안전운전(4). 한국레미콘공업협회. 레미콘 no.9 = no.25, 78-83.
5. 이경준, 정임국, 노윤환, 윤상경, 조영석. (2015). 도로위의 기상요인이 교통사고에 미치는 영향. 한국데이터정보과학회지, 26(3), 661-668.
6. 김윤용, 조길호, 김용구. (2020). 대구지역 교통사고 유형별 위험요인 분석. 한국데이터정보과학회지, 31(3), 503-510.
7. 최새로나, 이기영, 오철, 김동균.(2012).기상 및 교통조건이 고속도로 교통사고 심각도에 미치는 영향분석.대한교통학회 학술대회지,66(),247-252.
8. Yannis, Karlaftis, (2010). Weather Effects on Daily Traffic Accidents and Fatalities: A Time Series Count Data Approach. In Proceedings of the Transportation Research Board 89th Annual Meeting Compendium of Papers, Washington, DC, USA, 10-14.
9. 현대해상화재보험 교통기후환경연구소. (2021). 혹서기 교통사고 특성분석. 2021. 8. 5. 수정, 2023. 02. 15. 접속, https://blog.hi.co.kr/2480.
10. Meraldo Antonio(2019). Live Prediction of Traffic Accident Risks Using Machine Learning and Google Maps, 2023.02.12. 접속, Live Prediction of Traffic Accident Risks Using Machine Learning and Google Maps | by Meraldo Antonio | Towards Data Science
11. Kumar and Toshniwal (2015). A data mining framework to analyze road accident data. 2023.02.13. 접속, https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0035-y
12. 김영덕, 박광열, 이종관 and 최진영. (2007). 주행차량의 횡풍 영향에 관한 실험적 연구. 한국풍공학회지, 11(2), 161-170.

- 사용 데이터

지역별상세관측자료 (https://www.weather.go.kr/plus/land/current/aws_table_popup.jsp)
통계청 통계지리정보서비스, 대한민국 행정동 경계(admdongkor) (https://github.com/vuski/admdongkor)
행정안전부, (도로명주소) 도로구간 (http://data.nsdi.go.kr/dataset/12902)