지역 공공데이터 기반 범죄 발생 위험지역 예측 모델

- 전주시를 중심으로 -

Crime risk area prediction model based on local public data: Focus on Jeonju

김동완(Dongwan Kim) | 전북대학교 사회학과 학사과정 | asdf134652@gmail.com
이승재(Seungjae Lee) | 전북대학교 산업정보시스템공학과 학사과정 | grasianos10@naver.com

목 차

1. 서론
2. 이론적 배경
3. 연구방법
4. 연구결과
5. 결론

초 록

도시의 성장 및 경제적 수준 향상 등으로 인해 사회는 끊임없이 변하며, 위험사회가 도래하게 되었고 지역사회 불안요인에 범죄는 항상 언급되고 있다. 대부분의 범죄는 무작위적인 분포로 발생하지 않고 공간 및 지역에 따라 일정한 패턴을 보이며 발생한다. 이에 본 연구에서는 전주시의 공공데이터와 범죄 주의 구간 데이터를 활용해 범죄 발생구역의 공간적 특징들을 분석하고, 이에 따른 범죄 발생 위험구역을 기계학습 예측 모델을 통해 구축하고자 하였다. 분석 결과 분류 정확도 71.4%의 랜덤포레스트 모델을 구축하고, 그에 따른 변수 중요도를 산출했으며, 범죄 발생 위험구역을 예측하였다.

* 키워드 : 지역 공공데이터, 공간 빅데이터, 범죄분석, 기계학습, 예측 모델

ABSTRACT

Urban growth and economic level due to risks, and is constantly changing society has arrived, and crime is always mentioned in local factors of social unrest.Crimes are mostly random distribution in space and without a consistent pattern according to the region, occur.As a result, this study, Jeonju-si, in the public data and spatial characteristics of the crime zone by taking advantage of the crime care segment data analysis and, according to the crime to deployment with the prediction model through the danger zone. Classification accuracy analysis results, 71.4 percent of the random forests - calculation, and the crime of the variable importance according to build up the model, and then predicted the danger zone.

* KeyWords : Spatial Big Data, Crime Analysis, Machine Learning, Prediction Model

* 본 논문은 2019년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2019S1A5B8099507).
• 논문 접수 : 2020년 월 일 • 최초심사일: 2020년 월 일 • 심사완료일: 2020년 월 일

1. 서론

2019년 전북일보에 따르면 전북에서 4대 범죄(살인·강도·절도·폭력)가 가장 많이 발생한 곳은 전주 시 덕진구로 나타났다. 국회 행정 안전 위원회는 경찰청으로부터 제출받은 '전국 관서별 4대 범죄 발생 현황'자료에 따르면 2018부터 2019년 9월까지 전북에서 발생한 4대 범죄는 모두 1만 4447건이다. 시·군별로는 전주 시 덕진구가 2713건으로 가장 많았고, 익산 시 2670건, 군산 시 2629건, 전주 시 완산 구 2566건, 정읍 시 816건, 남원 시 574건, 완주군 509건, 부안군 417건, 고창군 352건, 무주군 138건, 순창군 121건, 진안군 120건, 임실군 104건, 장수군 88건 등 순으로 집계됐다. 살인과 절도 발생도 전주 시 덕진구에서 각각 8건과 1178건으로 가장 많았다. 강도는 전주 시 완산 구 7건, 폭력은 익산시가 1682건으로 도내에서 가장 빈번했다. 이렇듯 전주 시에서는 전라북도에서 높은 범죄 발생 빈도를 보이고 있어 범죄 구간을 예측하여 예방하는 것이 필요하다.(전북일보 인터넷신문, 2019.10.06., https://www.jjan.kr/news/articleView.html?idxno=2063459)

한편 2019년 경찰청 통계연보에 따르면, 우리나라 총 범죄 발생 빈도는 계속 감소하는 추세를 보인다고 하지만 '국가종합지표체계'의 총인구 대비 범죄 발생 건수로 계산한 형법 범죄율은 증가 추세이고 시민들의 범죄에 대한 두려움 역시 전체 20%로 적지 않다. 특히 5대 형법 범죄 중, 2000년에서 2016년 사이 형법 범죄의 주요 범죄인 절도는 1.1배, 성폭력은 3.9배, 폭행은 9.4배가 증가하였다고 한다.(통계청, 2017, www.kostat.go.kr)

5대 범죄 이외에도 21세기 들어 새롭게 등장하며 성장하고 있는 범죄는 사이버 범죄의 빅 데이터이다. 실제로 최근 사례에서 데이터 병합을 통해 피싱 공격에 사용하는 사이버 위협의 정황을 쉽게 발견할 수 있다. 이름, 이메일, 관심사, 비밀번호, 주소 등 해커가 훔친 데이터를 하나씩 병합하여 공격 대상에 대한 데이터베이스를 확장하며 피싱 공격에 사용하거나 신용 기록을 얻어 내는 것이다. 이렇게 빅 데이터로 구축된 데이터 세트는 기업의 조직 구조를 구체화하고, 해당 기업의 직원과 소비자에 대한 표적화를 가능하게 하여 공격을 더욱더 정교하게 만들고 있다.(펜타시큐리티시스템, 2021.6.4.)

이는 여전히 만연한 범죄에 대한 관련 요인 탐색 및 예측 요인 등을 분석하여 예방하는 것이 필요함을 나타낸다. 그러나 한국의 상황은 외국과 달리 범죄 데이터가 기본적으로 비공개이고 때문에 학술적 연구가 제대로 이루어지지 못하고 있는 문제점이 지적되고 있다. 그뿐만 아니라 범죄 자료를 지리 정보로 구축하지 못함으로써 고도의 범죄분석과 현장 중심의 수사 방범 정책에 마련에 어려움이 있다. 우리나라는 아직 공간적으로 동이나 구 단위로 제공되는 집계 자료를 이용하고 있는데, 이는 선진사례와 상당한 수준의 차가 있다고 할 수 있다. (최민제, 노규성. 2016).

따라서 우리나라도 범죄 예측을 하여 그 발생 빈도를 저감시키는 것이 사회 안전에 의미가 더 크다고 할 수 있으며, 이를 위해 빅 데이터 분석 기술과 데이터 마이닝

을 결합하여 범죄 동향을 필요한 조건에 따라 예측하려는 연구가 더 많이 필요하다고 볼 수 있다. (김원. 2020)

본 연구에서는 범죄 발생에 영향을 미치는 도시 내 물리·환경적 공간 빅 데이터를 구축하고 유관 변인들을 탐색 및 검증함과 동시에, 기계학습 분석을 활용하여 전주시의 범죄 발생 위험지역에 대한 심층 분석을 하고자 하였다.

이에 연구 방법으로 기계학습을 통해 범죄 위험 지역에 영향을 미치는 요인들을 파악하고, 전주시 관련 공간 빅 데이터에 대한 예측 모델을 제시하는데 그 목적이 있다.

2. 이론적 배경

2.1. 선행 논문을 통한 범죄관련 유형 분석

- 5대 범죄와 물리적 환경 영향요인의 상관성 분석

신민규 외 1명을 연구에서는 공간데이터의 속성정보에 범죄 발생에 영향을 미치는 다양한 물리적 환경요인이 있으나, 도로 폭, 도로의 포장재 질, 시야각도, 주변 건물 정보를 독립변수 선정하여 회귀분석을 진행하였다. 연구결과 강도 ·강간 범죄는 좁은 도로 폭과 시선 사각지대가 높은 영향력을 가지는 요인으로 나타났다. 이 연구는 물리적 환경요인으로 모든 범죄 발생 현상을 설명하는 것에는 한계가 보였다.(심민규, 김의명. 2018)

- GIS 공간분석 기법을 이용한 범죄 취약지 추출

범죄 예측에 활용코자 GIS 공간분석 기법을 이용하여 범죄 취약지를 추출하였다. 범죄 취약지는 범죄 통계자료를 이용하여 장소와 용도지역별로 다르게 발생되는 범죄를 GIS의 핫스팟 분석(Hot Spot Analysis)과 역 거리 가중법(IDW)을 이용하여 추출하였다. 또한 셉테드 (CPTED)의 감시 요소인 CCTV, 가로등, 지구대, 파출소에 대해서 각각 감시 범위와 가중치를 산정하고 범죄취약지도와 중첩하여 4개 등급(안전, 주의, 경고, 위험)으로 표현된 셉테드 기반의 범죄 취약지도를 제작하였다. 앞서 밝힌 2건의 선행연구는 실제 범죄 발생 장소를 예측하는 데 있어 한계가 있다는 점과는 다르게 이 연구는 빅 데이터와 GIS를 활용하여 범죄예측 지역을 예상했다는 점에서 유사한 점이 있다. 하지만 본 연구는 전주시를 기준으로 연구 범위를 250M*250M의 세밀한 단위로 설정했다는 차별성이 있다. (박소랑, 박재국. 2018)

2.2. 선행 논문을 통한 범죄예측 모델 분석

- 베이지안 확률 기반 범죄 위험지역 예측 모델 개발

허선영 외 2명의 연구를 살펴보면 도시 공간 내 가로 폭, 평균 층수, 용적률, 1층 사용 용도(제2종 근린생활시설, 상업시설, 유흥시설, 주거시설)을 변수로 물리·환경적

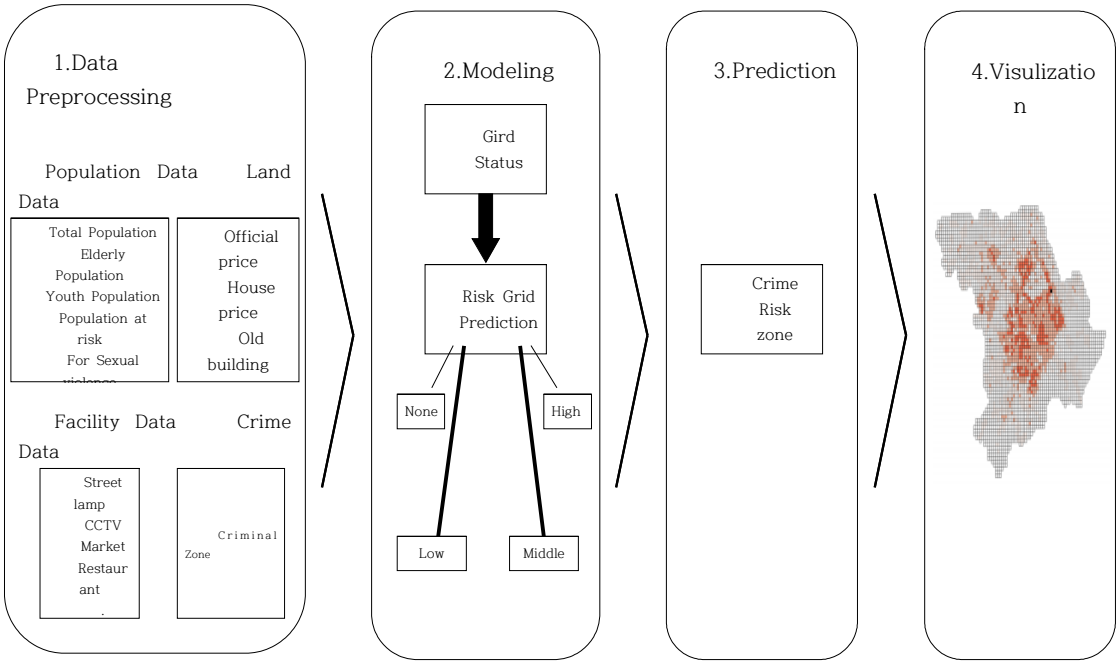
요소, 공간 빅 데이터를 구축 후 공간 회귀 분석을 실시하여 베이지안 모형으로 범죄 발생 위험성 예측 확률 모델을 개발하여 범죄 위험도를 분석을 하였다. 하지만 도시 사회 현상에 관한 많은 요인이 분명히 있을 것이나, 이를 공간 빅 데이터로 구축하지 못하여 적용하기 어려운 한계가 있다. (허선영, 김주영, 문태헌. 2017)

- 빅데이터 표준분석모델을 활용한 CCTV우선 설치지역 도출 사례연구

성창수 외 3명의 연구에서는 공공기관의 빅데이터를 활용한 표준 분석 모델을 가지고 적합성과 효과성을 확인을 하였는데 설치리스트 상위 지점 모두 범죄 취약지수 중 환경 지수 값이 전반적으로 낮게 나온 반면, CCTV 미설치 지역에 따른 감시 취약지수와 야간 및 심야 시간대의 유동인구 지수가 높게 나타났다. 이는 실제로 CCTV 설치에 대한 민원이 높고 그 필요성을 인지하고 있는 지역으로 빅 데이터 분석 결과가 높은 정확성을 나타내고 있음을 확인할 수 있다. 따라서 빅데이터 활용과 분석에 의한 표준 분석 모델의 의미 있는 시사점을 제시하고 있다. (성창수, 박주연, 가회광. 2017)

- 빅데이터 분석을 통한 순찰 및 112 신고 대응 효율화 모델

김중곤, 이태현 연구에서는 112신고에 보다 효율적으로 대처하기 위한 빅데이터 활용 모델을 제시하였다. SK텔레콤 유동인구 및 112신고 데이터 등을 활용하여 112신고 예측 모형을 추정한 후, 이를 바탕으로 법정동 별 112신고 발생 위험도를 예측한다. 둘째, 위험 예상 법정동 별로 112신고 군집화분석을 실시하여 예측 모형을 도출하였다. 하지만 예측 모형에 투입된 독립변수가 제한적이며 유흥업소나 CCTV 같은 변수들을 분석에 포함하지 못하였다. (김중곤, 이태현. 2020)



<그림 1> 범죄 위험지역 예측모델 도식화

3. 연구 방법

3.1. J시 범죄 위험지역 예측모델 개요

<그림 1>은 본 연구에서 목표로 하는 범죄 발생 위험지역 예측의 흐름을 도식화한 그림이다. 연구는 다음과 같이 크게 4단계로 구분된다. 먼저 전주시의 인구, 토지, 시설물, 범죄 주의 구간에 대한 데이터를 수집 및 병합하여 하나의 테이블을 구성한다. 그 후, 데이터 정제 과정을 거쳐 분석에 유의미한 데이터를 선별하고, 모델에 적합할 변수들을 선택했다. 해당 모델은 범죄 발생량을 4자기로 구분하여 위험도를 산출하며, 최종적으로 모델에서 범죄 발생량이 높을 것이라고 예측한 격자를 '범죄 위험지역'으로 전주시 지도에 시각화한다.

본 연구에서는 해당 전주시의 인구, 시설물, 지가 등 다양한 요인을 이용하여 모델을 구축하고자 한다. 이를 위해 먼저 국토정보 플랫폼을 이용하여 전주시의 격자별 인구, 노후 건축물, 공시지가 등의 데이터를 수집했으며, 공공데이터 포털에서 신호등, 상가, CCTV 등의 시설물 데이터를 수집하고 해당 시설물의 주소를 이용하여 지오 코딩을 통해 좌표를 추출했다. 종속변인으로 사용할 범죄 위치 데이터는 공개되지 않기 때문에 생활안전지도에서 제공하는 범죄 주의 구간 데이터로 대체하여 활용하였고, 격자 ID를 KEY로 데이터 통합 과정을 걸쳐 수집된 데이터를 하나의 데이터 셋으로 결합하였다.

본 연구에서 구축한 예측모델은 각 격자별 범죄발생량을 '없음', '낮음', '중간', '많음'으로 구분하여 예측하는 다중분류 모델로 진행하였으며, 데이터 분석과정에서 선택된 변수를 중심으로 기계학습 모델을 구현하였다.

본 연구에서 구현한 범죄 위험지역 예측모델은 다음과 같은 특징을 지닌다. 첫째, 250m x 250m로 설정된 세밀한 연구단위로 범죄발생에 영향요인을 파악한다. 둘째, 개별주택가격, 공시지가, 교통시설물 등 선행연구 대비 다양한 변수를 이용하여 예측모델을 구축한다.

3.2. 자료 수집 및 처리

본 연구는 인구수 약 60만 명이 거주하는 전라북도 전주시를 분석 대상으로 삼았다. 분석단위는 전주시의 행정동 전체를 하나의 유형으로 판단하기에는 동 내의 많은 요인들을 고려하기 어렵다고 판단하여, 250m x 250m 구역으로 분할했다. (곽명신, 서정주, 성형곤, 2017)

분석단위인 격자는 국토정보 플랫폼에서 250m x 250m 격자별 인구(총 인구, 유소년 및 고령인구 등), 공시지가, 개별주택 가격, 노후 건축물 등의 데이터를 수집하였으며, 선행연구를 참고하여 범죄 발생에 유의한 물리적 환경이라고 판단되는

CCTV, 가로등, 보안등 및 시설물 요인(주차장, 공중화장실, 버스정류장, 오피스텔 및 원룸, 음식점, 유흥 주점, 파출소 등)을 수집하였으며, 공공데이터 포털에서 제공하는 CSV 및 OPEN API를 이용하였다. (박소랑, 박재국, 2018)

데이터 병합 단계에서는 Q-GIS를 이용했는데, 250m x 250m 격자로 구성된 전주시의 인구 및 건축물 데이터를 격자 ID에 따라 '통합'기능을 이용해 결합하였으며, 전주시 행정동 지도를 이용해 격자별 행정동 코드를 부여하였다. 이후 시설물 관련 데이터 (상가, 가로등 등)의 주소를 이용해 지오 코딩을 진행하였고, 산출된 위도, 경도를 Q-GIS에 투영시킨 후 '폴리곤에 포함되는 포인트 개수' 기능을 이용해 격자별 시설물의 수를 계산하여 데이터에 입력시켰다.

범죄 위치 데이터는 국립재난안전 연구원에서 운영하는 생활안전지도 서비스에서 WFS-API를 통해 치안 정보 중 범죄 주의 구간의 전체 데이터를 추출하였다. 범죄 주의 구간 데이터는 5대 강력 범죄(강도, 성폭력, 폭력, 절도, 살인)발생 현황을 밀도 분석하여 도로상에 등급으로 부여한 정보이며, 1~10등급으로 표현되어 있다. 앞서 추출한 전주시 위치 데이터(250m x 250m)에 범죄 정보 폴리곤 데이터를 가장 높은 범죄 등급(max)으로 공간 조인했다.

GIS에서 생활안전지도의 치안 사고 통계를 기반으로 도로 레이어에 입력된 범죄 주의 등급 데이터를 추출한 결과 등급이 2~11등급이었다. 이에, 각 등급을 1단계씩 낮추어 1부터 10까지 맞추었고 격자 내 위험 등급이 없는 곳은 0값으로 지정했다.

<표 1>은 본 연구에서 활용한 기초 데이터를 정리한 것이며, 총 35개 변수와, 3,498개의 행으로 구성되어 있고, 122,430건의 데이터를 구축했다.

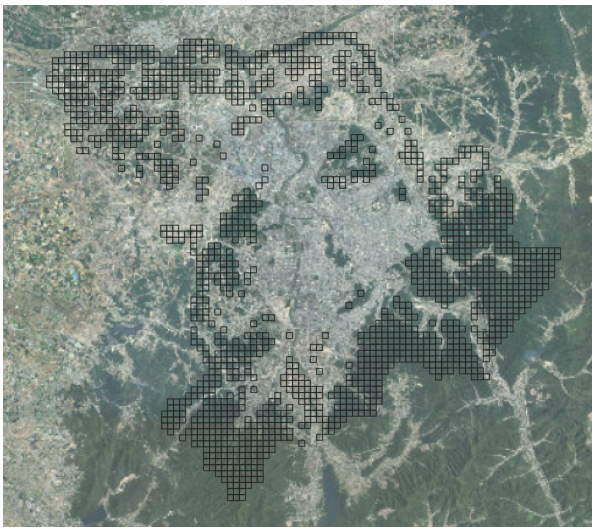
<표 1> 기초 데이터

데이터	내용	유형	출처	건수
생활안전지도	성폭력, 절도, 폭력 및 전체 범죄 주의 구간 등급	shp	행정안전부	3,498건
상가(상권 정보)	유흥주점, 음식점 등	csv	소상공인 진흥공단	20,988건
시설물 데이터	가로수, CCTV 등	csv, xml	전주시 스마트 시티과	62,964건
인구, 건축물 데이터	상주(주거)인구, 건축물, 공시지가 등	shp	국토정보플랫폼	34,980건

분석에 이용할 데이터에 대한 히스토그램을 탐색한 결과, 값이 0(NaN)인 경우가 많

았다. 이는 세밀한 분석을 위해 격자별로 구분하였기에, 각 격자에 포함되지 않는 변수들이 많은 것으로 보인다. 해당 격자들을 GIS에 투영해 확인해 본 결과 강이나 산간지역으로 확인되었다. 이에 이 해당 격자들은 이상치라고 판단하여 연구에서 제외하였다(그림 2).

데이터 처리 과정에서 ID로 사용될 독립된 격자는 총 3,498개였으며, 이 중 공시지가를 제외한 모든 변수의 값인 행을 탐색한 결과 1,460개의 행이 있었다. 본 분석에는 이를 제외한 2,038개의 격자만 이용하였다.



<그림 2> 전주시 제외 격자 위치

종속 변인으로는 전주시의 격자별 범죄 주의 구간 등급이었다. 독립 변인들에 대한 사전 데이터 탐색 결과 수집한 독립 변인(총 35개의 변인들)은 1,460개의 격자를 제거했음에도 대체로 왼쪽으로 치우친 분포를 띄었다. 이는 각 격자 간 시설물 및 인구의 분포의 편차가 크다는 의미이며, 전주시의 도심에 인구 및 인프라가 밀집되어 있음을 확인할 수 있었다. 또한, 각 변수마다 산점도를 그려 확인한 결과 이상치가 있는 격자들을 확인했으며, 해당 격자를 제거했다. 변수 선택을 위해 상관분석 및 분산분석을 실시하였으며, 데이터의 상관이 높은 것들 중 변수 간 다중 공선성이 높은 변수들은 분산분석을 통해 변수를 선택했고, 범죄 데이터와 관련성이 없다고 판단되는 것들을 제거하여 최종 15개의 독립 변인들을 추려낸 후, 데이터 정규화를 진행했다.(표 2).

데이터 정제 결과 총 30,555개의 고품질 데이터를 확보하였다. 이는 원시데이터의 약 25%에 해당하는 데이터였다.

<표 2> 분석 변수 설명

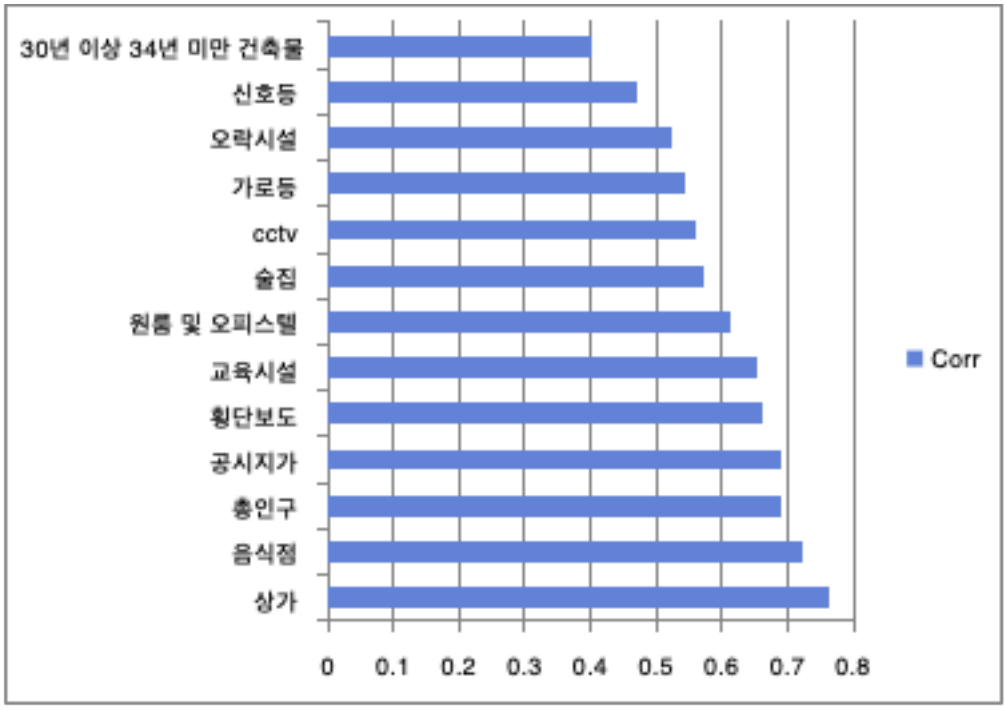
독립변수		종속변수(범죄 등급 구간)
인구	총 인구	범죄 주의 구간(0~10) (성폭력 + 폭력 + 절도 + 강도 + 살인)
토지	개별주택가격	
	공시지가	
	35년 이상 건축물	
시설물	공영주차장	
	버스정류장	
	오락시설	
	숙박업소	
	가로등	
	술집	
	원룸 및 오피스텔	
	음식점	
	횡단보도	
	신호등	
	cctv	

4. 연구결과

4.1. 변수 간 상관관계 분석

모델 적합에 앞서 수집된 데이터들의 변수 간 상관관계 분석을 진행하였다. 이는 범죄 발생에 관련성이 높은 변수를 선택하고, 그렇지 않거나 다중 공선성이 높은 변수들을 제거하기 위함이다. 상관관계 분석에는 종속변인이 서열형 변수일 때 사용하는 Spearman 상관분석을 진행하였으며, 종속변인인 범죄 주의 구간 등급을 중심으로 31개의 연속형 변수와의 상관관계를 살펴본 결과(그림 3), 모든 독립변수들이 범죄 주의 구간 등급과 양의 상관관계를 보였으며, 17개의 변수가 0.4 이상의 상관계수를 나타냈다. 상가와 음식점이 Spearman 상관계수 0.7 이상으로 종속변수와 가장 관련이 있게 나타났고, 다음은 총인구가 0.69로 높게 나타났다. 또한, 공시지가와 원룸 및 오피스텔도 유의한 상관관계를 나타낸 것으로 보아, 전주시 범죄는 대부분 도심지역에서 발생하는 것을 추측해볼 수 있었으며, 30년 이상 34년 미만 건축물에서도 유의한 상관관계를 보여 노후 건축물이 범죄 발생량과 연관이 있을 것이라 미루어볼 수 있었다.

하지만 해당 독립변수들은 변수 간 상관관계가 높아 다중 공선성 문제가 발생할 수 있겠다고 판단되었고, 이를 해결하고자 연관성 있는 변수 중 종속변수를 가장 잘 설명할 수 있는 변수를 채택하고자 분산분석과 사후 검정을 실시하였다.



<그림 3> 범죄 주의 구간 등급을 기준으로 변수 간 상관계수

4.2. 변수 간 분산분석

상관성이 높은 독립변수와 종속변수에 대한 산점도를 살펴본 결과, 0~10으로 이루어진 종속변수가 유사한 분포를 띄고 있다는 사실을 발견했다. 이에, 독립변수에 대한 종속변수의 구간 별 평균 차이에 대한 검정을 실시한 결과, 유사한 분포를 띄는 구간의 범죄 발생량을 그룹화하였다. 종속변수 값이 0인 그룹은 '없음', 1~3인 그룹은 '적음', 4~7인 그룹은 '중간', 8~10인 그룹은 '많음'으로 구분하였다.

이후 다중 공선성 제거를 위해 분산분석 및 Bonferroni 사후 검정을 실시하였는데, 변수별 총 4개 그룹 6개의 비교 중 4가지 이상의 비교가 유의미한 변수를 채택하고자 하였으며, 비교가 유의미함에도 변수 간 상관관계가 큰 경우에는 p-value가 낮은 변수를 선택하거나, 종속변수와 가장 상관관계가 높은 변수를 채택하며 여러 방식으로 모델에 적합하고자 하였다.

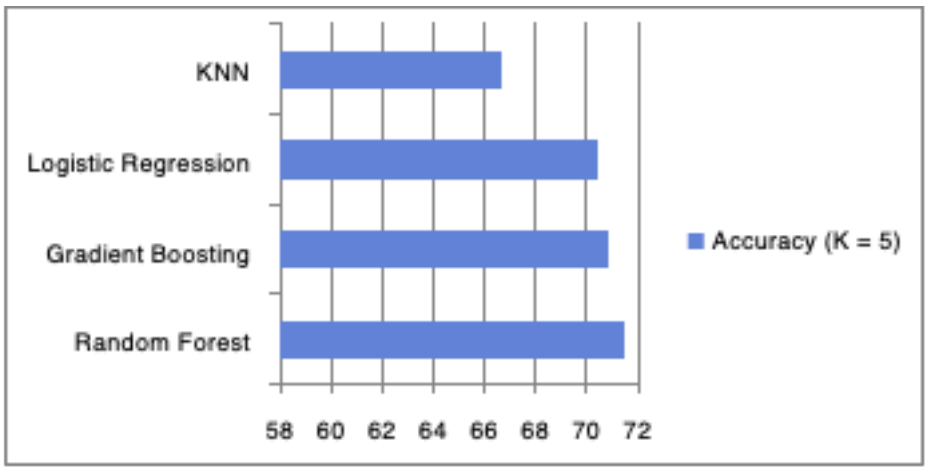
분산분석 결과 채택된 변수는 총인구, 개별주택 가격, 공시지가, 공영주차장, 버스정류장, 오락시설, 35년 이상 건축물, 숙박업소, 술집, 원룸 및 오피스텔, 음식점, 가로등, 횡단보도, 신호등, CCTV로 총 15개의 변수를 채택하였다. 해당 변수들의 분산팽창 인수(VIF)로 다중 공선성을 확인해 보았을 때 모두 3 미만으로 변수 간 다중 공선성 문제를 해결할 수 있었다.

4.3. 모델 학습

모델 학습에 앞서 해당 데이터 셋은 전주시의 모든 격자에 대한 범죄 위험지역의 예측을 목적으로 하기에, K-fold 데이터 분류를 실시하였다. 이에 K 값을 5로 설정하였으며, Train set 80%와 Test set (20%)로 5번의 검증을 실시하고 각 Fold 별 Accuracy의 평균으로 모델의 성능을 평가하고자 하였다. Training set은 약 1,630개의 데이터를 이용했고, Test set은 약 407개의 데이터를 이용하였다.

기계학습 모델 적합 시 Random Forest, Gradient Boosting, Logistic Regression, KNN 모델을 이용하였으며, Random Forest 모델이 평균 Accuracy 71.4%로 가장 높은 성능을 보였다. (그림 4)

해당 모델에 대한 분류 성능 평가 지표를(표 3) 살펴본 결과, 해당 모델은 범죄가 발생한 지역과 발생하지 않은 지역에 대한 분류 정확도는 약 90%로 높은 수준을 보였으나, 범죄가 많이 발생한 지역(범죄 주의 구간 등급 8~10)에 대한 예측력은 높지 않은 한계를 보였다.



<그림 4> 모델별 정확도

<표 3> 분류 성능 지표

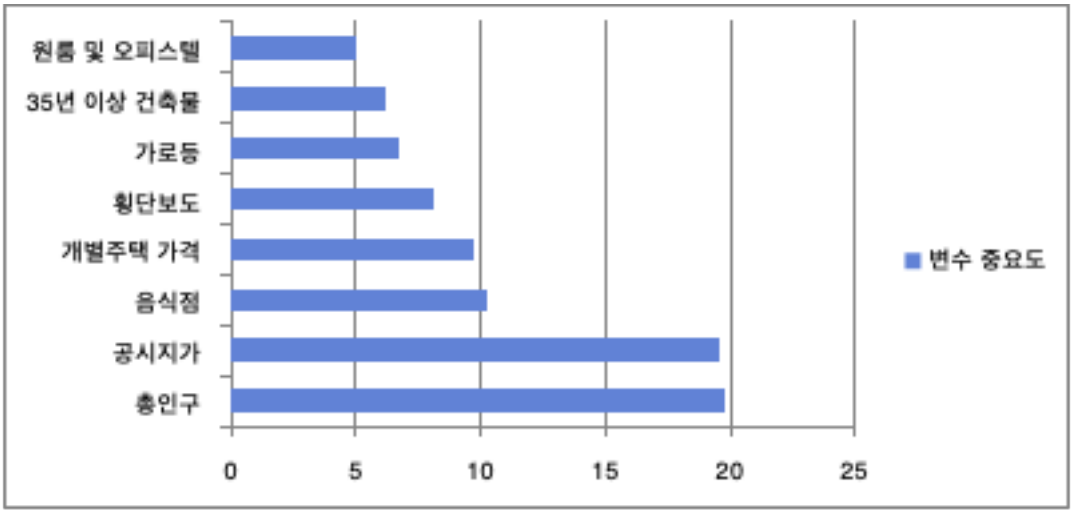
		Predict			
		없음	적음	중간	많음
Actual	없음	910	110	2	0
	적음	236	402	76	0
	중간	8	132	142	0
	많음	0	6	12	1

연구의 첫 번째 목적인, 범죄 발생에 미치는 영향 요인을 파악하기 위해 가장 성능이 좋았던 Random Forest모델에 Train set 70%와 Test set (30%)로 데이터를

적합한 결과, 모델의 정확도는 74.3%였으며 본 분석에서 이용한 교차 분석보다 더 높은 정확도를 보였다.

해당 모델의 변수 중요도를 수치화 해본 결과 총인구, 공시지가, 음식점, 개별주택가격, 횡단보도, 가로등, 35년 이상 건축물, 원룸 및 오피스텔 등이 그 뒤를 이었다.(그림 5)

이는 해당 모델이 전주시의 범죄 위험 지역을 예측할 시, 격자 별 거주하는 인구가 가장 큰 영향을 미쳤다는 것을 알 수 있으며, 토지 및 건축물의 가격 또한 큰 영향을 보인다는 것을 알 수 있다.



<그림 5> 랜덤 포레스트 모델의 범죄 주의구간 분류에 따른 변수 중요도

5. 결론

현시대는 인공지능 기술 활용을 확대해 데이터 기반의 과학적 보호관찰을 활성화고 첨단 기술을 융합한 효과적인 강력 범죄 재범 예방 체계를 강화해 사회안전망을 보다 촘촘히 하는데 노력하고 있다. 이를 토대로 본 연구는 전북 4대 범죄가 가장 많이 발생하는 전주의 공간 데이터 분석을 통해서 범죄 위험 지역 예측을 진행하였다.

본 연구는 전주시의 범죄 발생 위험지역을 예측하는 모델을 구축하였다. 국토정보 플랫폼, 공공데이터 포털, 생활안전지도를 통해 데이터를 수집하였으며, 데이터 병합 단계로 Q-GIS를 이용하여 총 35개 변수와, 3,498격자로 구성된 122,430건의 기초 데이터를 얻었다. 변수의 기초통계량을 확인 후 산간지역의 격자를 제거하였으며, 히스토그램을 탐색한 결과 발견된 이상치 또한 제거 후 상관분석 및 분산분석을 실시하였고, 최종적으로 30,555개의 데이터를 확보하였다. 데이터 정규화 후 Random Forest 모델을 활용하여 예측 모델을 평가한 결과 정확도는 71.4%였으며, 외곽지역은 다소 예측률이 떨어졌지만 도심의 범죄 발생 현황을 유사하게 예측함을 볼 수 있

었다. 범죄 발생량 '많음' 등급의 격자는 전주시 우아동의 전주역 부근으로 1곳만이 정확히 예측이 되었다. 해당 지역은 인구수는 높은 수준이 아니었으나, 공시지가가 3사분위수를 초과할만큼 높은 수준이었으며, 유흥시설인 술집, 숙박업소들 및 음식점이 많이 분포되어 있었다. 하지만 CCTV의 개수는 1개로 현저히 부족한 수준이었고, 주변 격자 또한 위험격자로 분류되어있어, 해당 구역에 대한 CCTV 확충 및 순찰 강화가 필요하다고 판단되었다.

본 연구는 선행연구에서 범죄 위치 데이터는 개인정보 보호법상 비공개 처리가 많아 학술적 연구가 제대로 이루어지지 못하고 있는 문제점을 극복하고자 범죄와 관련된 다양한 요소의 데이터를 수집 및 분석을 하였고, 연구 범위를 250m x 250m의 세밀한 단위로 설정했다는 차별성이 있다. 또한, 수집이 까다로운 범죄 발생 위치 데이터를 대체하여, 생활안전지도 서비스에서 WFS-API를 통해 범죄 주의 구간 데이터를 수집하여 분석한 결과 외곽 지역을 제외한 높은 예측률을 확인한 점에서 의미 있는 결과를 도출하였다는 의의가 있다. 데이터 정제 측면에서는 원시데이터를 정제 후 약 25%에 해당하는 데이터를 사용하여 모델을 구축한 결과, 다량의 데이터 보다 고품질의 데이터 선별을 하는 과정의 중요성이 돋보였다.

본 연구의 한계점으로는 데이터 수집 과정에서 유의한 요인이라 판단되었던 유동인구, 격자별 1인 가구 거주 비율 등 데이터를 수집하는 것에 한계가 있었으며, 종속변인을 격자별 범죄주의구간 등급의 최댓값으로 공간 조인하여 미세한 구간 차이로 데이터가 오 분류되는 문제가 있었다. 또한, 범죄 발생률이 많은 격자에 대한 예측력이 높지 않았다.

본 연구의 기대효과로 첫째, 향후 관공서와 협의를 통해 범죄에 유관한 데이터 및 데이터를 구축할 경우 설명력 높은 모델을 구현할 수 있을 것이며 둘째, 모델 성능 개선을 통해 범죄 발생 구역 예측에 대한 표준 분석 모델 개발의 사전 연구로 활용될 것이다.

참고문헌

허선영, 김주영, 문태헌. (2017) 베이지안 확률 기반 범죄위험지역 예측 모델 개발(한국지리정보학회지), 20(4), 89-101.

최민제, 노규성. (2016). 빅데이터 융합 기반 범죄예방에 관한 탐색적 연구 - 성남시 사례분석을 통해(디지털융복합연구), 14(11), 125-133.

김신혜, 김광열, 백태경. (2021). 셉테드(CPTED)를 이용한 도시재생방안에 관한 연구 - 부산시 안심마을 조성사업 대상지를 중심으로(한국지리정보학회지),24(1), 54-67.

신민규, 김의명. (2018), 5대 범죄와 물리적 환경 영향요인의 상관성 분석(한국지도학회지), 18(3), 131-140.

박준휘 외 7 (2014). 범죄유발 지역·공간에 대한 위험성 평가도구 개발·적용 및 정책대안에 관한 연구(Ⅱ), 형사정책연구원 연구총서 . 1-608.

박소랑, 박재국 (2018) 공간 빅데이터와 범죄통계자료를 이용한 범죄취약지 추출 (융합정보논문지) 8.1, 161-171.

허선영, 문태헌. (2013). 범죄다발지역의 도시 환경적 영향요인 분석.(국토계획) 48.6, 223-234.

허선영 , 김주영, 문태훈. (2018). 머신러닝기반 범죄발생 위험지역 예측. (한국지리정보학회지), 21(4), 64-80.

박현수. (2018). 범죄 두려움에 영향을 미치는 요인의 공간 분석. (형사정책연구) 29.2 91-117.

박지호 . (2015). 범죄발생 요인 분석 기반 범죄예측 알고리즘 구현. (한국위성정보통신학회 논문지), 10(2), 40-45.

성창수, 박주연, 가회광. (2017). 빅데이터 표준분석모델을 활용한 CCTV우선 설치지역 도출 사례연구. (디지털융복합연구), 15(5), 61-69.

김중곤, 이태현. (2020). 빅데이터 분석을 통한 순찰 및 112 신고 대응 효율화 모델: 대구시의 사례를 중심으로. (한국치안행정논집), 17(3), 91-106.

김원. (2020). 빅데이터 처리 기반의 범죄 예방 스마트 시스템에 관한 연구. (한국융합학회논문지), 11(11), 75-80.

안찬식, 오상엽. (2010). 스피어만 상관계수를 이용한 사용자 상황 및 특성 처리 개선.(멀티미디어학회 논문지), 13(19) 1444-1452

검찰청 (2020) 전라북도 전주시 5대 범죄의 통계자료. 검색일자: 2021. 08. 5. www.spo.go.kr

통계청 (2017) 장래인구추계에 대한 통계자료. 검색일자 : 2021. 8. 3 www.kostat.go.kr

한국형사정책연구원(2018) 국민생활안전실태조사 실시 : 2021. 8. 1 www.kic.re.kr

국한문 참고문헌의 영문 표기

(English translation / Romanization of reference originally written in Korean)

Heo-Sunyoung, Kim-Jooyoung .Moon-Taeheon (2017) Crime Incident Prediction Model

based on Bayesian Probabilityl. The Korean Association of Geographic Information Studies, 20(4), 89-101.

Min-Je Choi, Noh,Kyoo-Sung. (2016). Exploratory Study on Crime Prevention based on Bigdata Convergence - Through Case Studies of Seongnam City. Journal of Digital Convergence., 14(11), 125-133.

Kim-Shinhye. Kim-KwangYeol. Baek-Taekyung. (2021). A Study on Urban Regeneration Considering the CPTED - Focusing on the Case Study of the Busan Ansim Village Project. The Korean Association of Geographic Information Studies, 24(1), 54-67.

Shin Mingyu. Kim Eui Myung.. (2018), Analysis of Relation between Five Crime Types and Physical Environmental Factors. The Korean Association of Geographic Information Studies, 18(3), 131-140.

Park-Junhwi, Kang-Yonggil. Kim-Dowoo. (2014). The Development of Crime Risk Assessment Tool and Its Application in South Korea(Ⅱ)- with focus on commercial are, Korean Institute of Criminology and Justice. 1-608.

Heo-Sunyoung, Moon-Taeheon (2013). Analysis of Urban Environmental Impact Factors in Crime Hotspot. National Land Planning. 48.6 223-234.

Heo-Sunyoung, Kim-Jooyoung .Moon-Taeheon. (2018). Predicting Crime Risky Area Using Machine Learning. The Korean Association of Geographic Information Studies, 21(4), 64-80.

Park-Hyunsoo. (2018). Spatial Analysis of Factors Affecting Fear of Crime. Korean Institute of Criminology and Justice. 29.2 91-117.

Park-Jiho, Cha-Kyunghyun. Kim-kyung ho, Lee-Dongchang, Son-Gijun, Kim-Jinyoung. (2015). Implementation of Crime Prediction Algorithm based on Crime Influential Factors. Korea Society of Satellite Technology, 10(2), 40-45

So-Rang Park, Jae-Kook Park (2018) Extraction of Crime Vulnerable Areas Using Crime Statistics and Spatial Big Data (Convergence Society For SMB) 8.1 161-171.

Chang Soo Sung,Joo Y. Park, Hoi Kwang Ka. (2017). The Case Study of CCTV Priority Installation Using BigData Standard Analysis Model. (Journal of Digital Convergence), 15(5), 61-69.

Kim-Joonggon , Lee Tae-Hyun. (2020). Big Data Analyses Models for Efficient Patrol and 112 Crime Call Response: Focusing on the Case of Daegu Metropolitan City. (Journal of Korean Public Police and Security Studies), 17(3), 91-106.

Won Kim. (2020). A Study on the Crime Prevention Smart System Based on Big Data Processing. (Korea Converence Society), 11(11), 75-80.

Chan-Shik Ann, Sang-Yeob Oh. (2010). Improvement of User's Context Aware and Characteristic Process using spearman correlation coefficients. (JOURNAL OF KOREA MULTIMEDIA SOCIETY), 13(19), 1444-1452.

Kucuk U, Eyuboglu M, Kucuk HO, Degirmencioglu G. (2016). Importance of using proper post hoc test with ANOVA. Int (International Journal of Cardiology), 10.1016, 209-347.

Brantingham, P. J., & Brantingham, P. L. (Eds.). (1981). Environmental criminology (pp. 27-54). Beverly Hills, CA: Sage Publications.

Prosecution service Korea (2020) Statistical data of 5 major crimes in Jeonju, Jeollabuk-do. Search date: 2021. 08. 5 www.spo.go.kr

Statiscics Korea (2017) Statistical data for future population projections. Search date : 2021. 8. 3 www.kostat.go.kr

Korean Institute of Criminology and Justice(2018) Conducted a National Living Safety Survey : 2021. 8. 1 www.kic.re.kr