**빅데이터 분석 결과 보고서**

**“치안 정보 분석 및**

**예측 서비스”**

**New Seekers**

송양민, 장효선, 안승주

**목차**

1. **분석 배경 및 기획 의도**
   1. 분석 배경
   2. 기획 의도
2. **분석 목표**
3. **분석 데이터**
   1. 데이터 수집
   2. 데이터 정합성 검정
   3. 데이터 전처리
4. **분석 과정 및 결과**
   1. 범죄 비율 분석 결과
5. **활용 방안**
6. **참고 문헌**
7. **분석 배경 및 기획 의도**
   1. **분석 배경**

2023년 7월 21일 신림역 칼부림 사건, 8월 3일 서현역 칼부림 사건, 그리고 동 시기에 발생한 470여건의 온라인 살인예고 등 불특정 다수를 대상으로 한 범죄들은 대한민국을 떨게 만들었다. 길거리에서 흉기를 소지한 사람이 있다는 내용의 뉴스가 1시간마다 보도될 정도였으며, 기존의 범행 유형이 특정인을 목표로 치정이나 원한관계 등에 의한 보복적 성격을 가지던 것과 달리, 범행 대상이 불특정 다수였기에 주요 쇼핑몰에서의 호신용품 매출이 전년 동기 대비 최대 460%까지 증가하는 등 시민들의 불안감이 커졌었다.

경찰의 발 빠른 초강경 대응으로 추가적인 혼란 없이 사태는 진정되어 갔으나 시민들의 불안감은 쉽게 가라앉지 않았는데 그 이유는 실제적인 범죄 발생이 불안의 원인이 아니라 ‘묻지 마’ 용어 등 SNS등을 통해 확대 재생산되는 단순 흥미위주의 자극적인 정보들에 원인이 있다고 전문가들은 분석했다.

이에 실제 치안 요소들과 범죄 발생 간의 관계를 분석하고 현재 치안 수준을 가늠하며 미래의 치안 수준을 예측하고자 한다.

* 1. **기획의도**

거주지역의 치안 시설물 현황에 기반한 범죄예측 서비스를 개발하여 시민들에게 현재 치안수준이 어느 정도의 범죄 예방 효과를 보이고 있는지 나타내고 불안감을 해소하고자 한다.

CCTV나 방범등과 같은 치안시설물들에 관한 통계자료와 주요 5대 범죄 발생에 관한 통계자료, 그리고 범죄유형에 따라 지역적 성격을 나타낼 수 있는 인구 및 특정 업종유형에 관한 통계자료들을 분석하여 요소들 간의 관계를 밝히고 예측하여 미래 치안 수준에 대한 가늠을 제공하여 시민들이 보다 정확한 현황을 파악함으로써 무차별적으로 재생산되는 막연한 불안감을 해소시키는데 그 목적을 둔다.

1. **분석 목표**

* 범죄율과 치안 시설물에 관한 공공데이터를 분석하여 주요 5대 범죄 유형과 관련된 요인들을 선별 및 특정한다.
* 지역적 특성을 반영할 수 있는 예측모델의 개발을 통해 보다 유의미한 정보를 제공함을 목표로 한다.

1. **분석 데이터**
   1. **데이터 수집**

- 치안시설 데이터: CCTV, 경찰관서, 방범등 \_서울 열린 데이터 광장, 공공 데이터 포털, 국가 통계 포털, 스마트치안 빅데이터 플랫폼

> 사전 조사에서 파악한 바, 치안 요소들은 CCTV와 같이 위치가 고정된 유형적 성격인 것들과 ‘심야 귀갓길 동행 서비스’와 같이 무형의 서비스 형태를 띄는 요소들로 나누어 볼 수 있었다. 무형의 서비스 형태인 요소들은 정량적 계측이 어려우며 제공되는 지역의 제한이 있음을 고려하여, 유형적 성격의 요소들을 주요 분석 대상으로 삼았다.

- 범죄 데이터: 주요 5대 범죄 발생률, 검거 율 \_서울 열린 데이터 광장, 국가 통계 포털, 공공 데이터 포털

> 주요 분석 대상 데이터인 만큼 정확성 높은 자료를 찾고자 하였다. 범죄의 발생과 검거의 경우 경찰관서들이 기준이 되는데 관할구역이 소재구의 경계와 일치하지만은 않아 경찰관서의 소재지에 따른 재 분류를 실시, 자치구별 데이터로 조사하였다.

- 지역 데이터: 지역 인구 수, 인구밀도, 면적, 유흥주점 수, 1인 가구수, 지역내 총생산\_서울 열린 데이터 광장, 공공 데이터 포털, 국가 통계 포털, 스마트치안 빅데이터 플랫폼  
> 범죄와 연관이 있을 것으로 예상되는 많은 요인들 중 지역적 성격을 나타낼 수 있는 요소들을 선정하는데 집중하였다. 인구 수는 인구 총 조사를 통해 안정적인 데이터를 얻을 수 있었으나 유흥주점의 경우, 업종이 일반음식점인 곳에서도 주류 판매가 가능하기에 이를 토대로 자료를 선정하면 지역적 성격과 분석 목표가 왜곡될 우려가 있어 보다 국소적이지만 분석목표를 표상할 수 있을 것으로 예상되는 유흥주점의 데이터를 조사하였다.

* 1. **데이터 정합성 검증(데이터 결손, 중복, 이상 확인 및 처리)**

데이터들의 공통적인 목표는 2004년부터 2022년까지의 자료를 수집하는 것이다.

- 서울시 5대 범죄 발생현황 통계\_서울 열린 데이터 광장

2014년부터 2022년까지의 자치구별 주요 5대 범죄의 발생 및 검거 데이터 수록

2004년부터 2013년까지는 서울시의 주요 5대 범죄 발생 및 검거 데이터만 수록

따라서 2004년부터 2013년까지의 자료는 “공공데이터 포털”의 “경찰청\_서울특별시지방경찰청\_관서 별 5대범죄 발생 및 검거 현황\_#2004~2013”을 이용하였다.

>공공데이터 포털의 데이터를 취합한 결과 2004년부터 2013년까지의 총 합계가 서울 열린 데이터 광장의 값과 달랐다. 이는 경찰 관서 별 관할 구역에 따른 차이로 보여 졌으며 관서 소재지를 기준으로 취합 자치구별 데이터로 재 분류하였다. 따라서 세부 수치가 근거가 되는 공공데이터 포털의 데이터로 갈음하였다.

>범죄 발생 수보다 검거 수가 많은 경우가 존재하였다. 오차가 있는 데이터를 삭제하는 방식을 택할 경우 데이터 손실이 23.8%(252행 중 60행) 예상되어 오차 값 정정을 실시하였다.

기준은 검거수가 발생수보다 작을 경우 원 값을 유지하였으며 반대의 경우 발생수로 갈음하였다.

- CCTV\_서울 열린 데이터 광장

범죄 관련 핵심 요인이지만 데이터가 2015년부터 2023년치까지만 존재했다.

2011년 8월 26일자 CCTV뉴스 『서울 중구, 주민 안전 위해 방범 용 CCTV 추가설치』 기사에서 2004년부터 2009년까지의 설치 수를 찾을 수 있었고,

2006년 5월 9일자 동아일보 『강남구”방범 용 CCTV설치중단”』기사에서 마찬가지로 2004년부터 2006년까지의 설치 수를 확인할 수 있었다.

CCTV 설치 수는 지속적인 증가세였기에 사이 값을 R\_ZOO PAKAGE의 na.approx()기능을 이용 선형보간을 실시하였다.

-서울시 도로시설물 통계(보안등)\_서울 열린 데이터 광장 및 서울시 시내주요기관 (경찰·소방관서) 통계 - 서울 열린 데이터 광장 에서는 결측 값이 발견되었으나 전체 데이터에서 차지하는 비율이 미미하여 삭제하였다.

- 서울시 인구밀도 (구별)통계\_서울 열린 데이터 광장

> 2007년까지의 자료만 있었으며 통계청의 인구 총 조사 자료 역시 2005년과 2010년만 존재하여 두 값을 합치고 선형보간으로 중간 값을 채웠다.

- 서울시 1인 가구 연령별 통계\_서울 열린 데이터 광장  
- 1인가구비율(시도/시/군/구)\_통계청(인구조사과)  
>두 데이터의 보충 결합을 위해 서울시 1인 가구 연령별 통계 값을 전체 인구수로 치환해 비율로 전환하였으며 사이 값을 선형보간으로 채웠다. 산식은 아래와 같다.

1인가구비율=(1인가구수÷일반 가구 수)×100

- 식품위생업 현황(구별\_유흥업소 수)\_서울 열린 데이터 광장

> 주류의 판매가 범죄에 유의미한 영향을 줄 것이라는 사전예측이 있어 주류판매를 하지 않는 휴게음식점의 수를 제외한 나머지 항목의 합산 값을 취합하였다.

- 자치구별 지역 내 총생산(2015년기준)\_서울 열린 데이터 광장

> 결측 값이나 이상 값은 없었으나 자료가 2010년까지만 존재하여 2004년부터의 자료는 연도별 증감 추이가 일정치 않았기에 평균값으로 대체하였다.

* 1. **전 처리**

지역내 총생산과 인구수, 보안등의 경우 값이 다른 데이터들과 범주 차이가 커 0과 1 사이 값으로 표준화를 진행하였다.

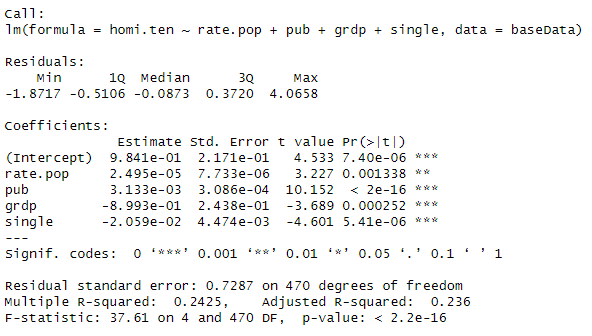
모든 데이터에서 사전 보간 처리 하지 않은 이외의 NA나 ‘-‘ 값은 삭제 처리하였다.

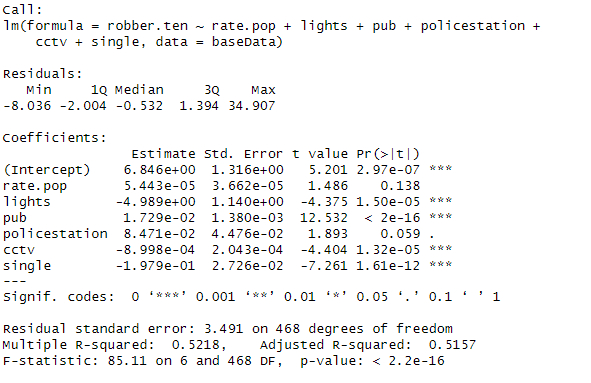
1. **분석 과정 및 결과**
   1. 범죄 비율 분석 결과 및 시각화
      1. 5개 범죄 분석
         1. 산점도를 통한 상관관계 파악

|  |
| --- |
|  |
| [표1-1] 범죄(살인)수와 변인들의 산점도 |
| 살인의 경우 뚜렷한 관계를 내는 변인들을 찾아보기가 힘들다. |
|  |
| [표1-2] 범죄(강도)수와 변인들의 산점도 |
| 강도의 경우 1인가구수, CCTV와 약한 음의 관계를 가지고 있는 것처럼 보이며 주점과는 약한 양의 관계를 가지고 있는 것으로 보인다. |
|  |
| [표1-3] 범죄(성범죄)수와 변인들의 산점도 |
| 성범죄의 경우 산점도 상으로는 지역내 총생산(GRDP)과 1인가구수, 방범등과 양의 관계를 보이며 주점과는 음의 관계를 보이는 것으로 분석된다. |
|  |
| [표1-4] 범죄(절도)수와 변인들의 산점도 |
| 절도의 경우 CCTV와 음의 관계를 가지는 것으로 보인다. |
|  |
| [표1-5] 범죄(폭력)수와 변인들의 산점도 |
| 폭력의 경우 꽤 뚜렷한 양상을 보여주는데 CCTV, 1인가구, GRDP와 CCTV, 그리고 1인 가구수와 음의 관계를 갖는 것으로 보이며 pub과 강한 양의 관계를 갖는 것으로 분석된다. |

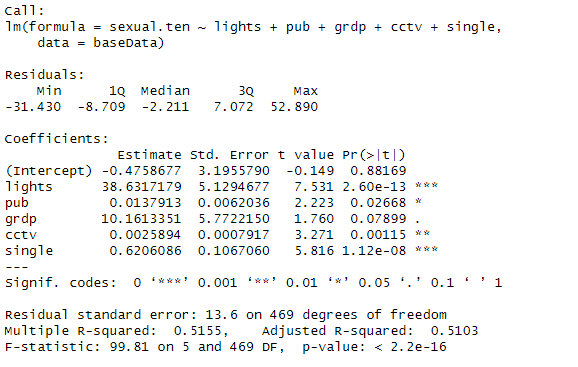
* + - 1. 회귀분석을 통한 변수 선정

1. 1차 회귀분석

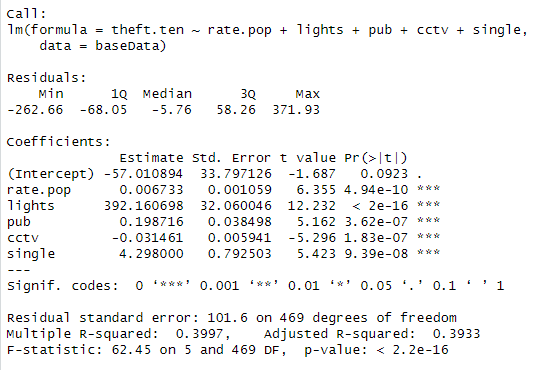
[그림1-1] 종속변수: 인구 10만명당 살인 수, 독립변수: 인구밀도, 유흥업소 수, 지역내 총 생산, 1인가구 비율



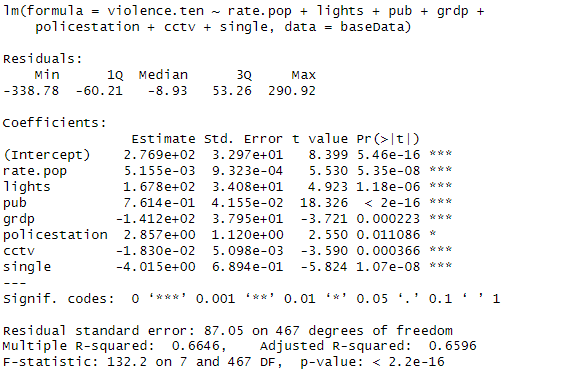
[그림1-2] 종속변수: 인구 10만명당 강도 수, 독립변수: 인구밀도, 보안등 수, 경찰서 수, CCTV수,1인가구 비율



[그림1-3] 종속변수: 인구 10만명당 성범죄 수, 독립변수: 보안등 수, 유흥업소 수, 지역내 총 생산, CCTV수, 1인가구 비율



[그림1-4] 종속변수: 인구 10만명당 절도 수, 독립변수: 인구밀도, 보안등 수, 유흥업소 수, CCTV 수, 1인 가구 비율



[그림1-5] 종속변수: 인구 10만명당 폭력 수, 독립변수: 인구밀도, 보안등 수, 유흥업소 수, 지역내 총 생산, 경찰서 수, CCTV 수, 1인가구 비율

상기 5개의 모델의 공통점은 높은 p-value값에도 불구하고 Adjusted R-squared 값이 매우 낮아 모델의 설명력이 충분치 않음을 알 수 있다.

이를 해결하기 위해서 단계적 접근을 실시하였다.

1. 특이 값을 R의 boxplot함수를 사용하여 검색하고 특이 값을 포함한 행의 삭제를 실시 후 다시 모델을 선정해 보았다

→ 오히려 Adjusted R-squared이 낮아지는 결과가 도출되었다. 이는 특이 값이 견인하고 있던 선형이 특이 값의 부재로 변형되면서 생긴 결과로 분석되었다.

1. 기존에는 자치구 전체 및 연도 전체를 분석 대상으로 삼았던 것을 각 연도별로 나누어 분석을 실시하였다.

→ Adjusted R-squared의 값 뿐만 아니라 p-value의 값도 낮아져 신뢰도가 더 낮아짐을 확인할 수 있었다. 이는 연도의 추이간 관계가 약하다는 것으로 분석되었다.

1. 자치구별로 나누어 분석을 실시하였다.

→ 가장 나은 결과를 보여줬다. Adjusted R-squared의 값은 0.5~0.8 사이에서 관측되었다. 즉 자치구의 지역적 특성이 반영되어 있는 독립 변수들인 만큼 자치구 기준에서 분석할 때 보다 적합함을 알 수 있었다. 하지만 아직 충분히 만족할만한 수치가 아니라 추가적인 보완 방법을 탐색하였다.

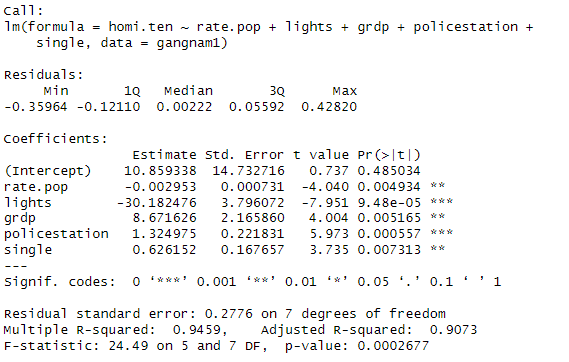
1. 기존에 계열 추세를 통해 보완하였던 자료들이 주로 포진해 있는 2004년~ 2009년의 데이터의 전 처리 과정을 재 실시하여 이번에는 결측 값을 포함한 행의 삭제를 실시한 후 회귀분석을 실시하였다.

→ 만족할만한 결과값이 도출되어, 기존 결측 값의 보완 방법이 적합하지 않았음을 알 수 있었다.

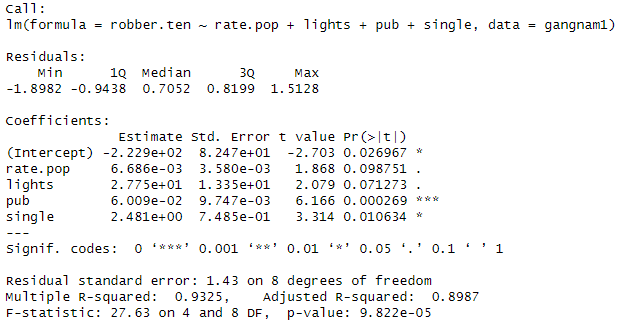
1. 2차 회귀분석

>각 구별로 나누어 진행

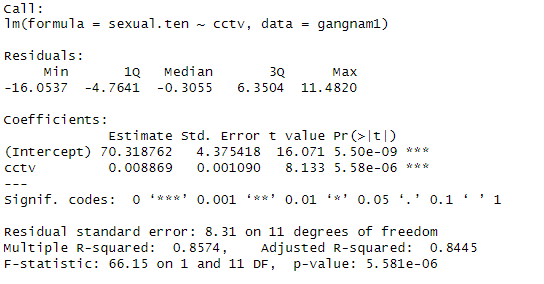
>강남구



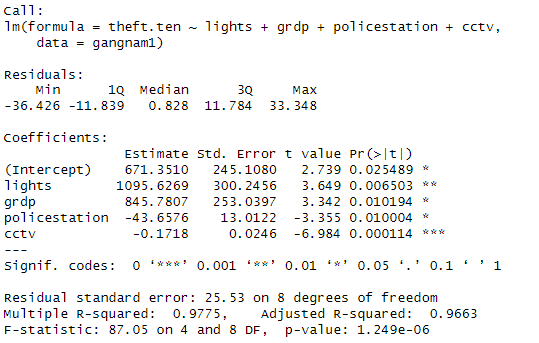
[그림2-1]전 처리 재 실시 후 살인 범죄 수 및 요인들



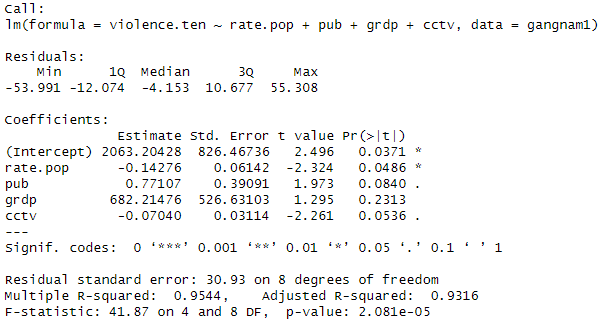
[그림2-2]전 처리 재 실시 후 강도 범죄 수 및 요인들



[그림2-3]전 처리 재 실시 후 성범죄 수 및 요인들

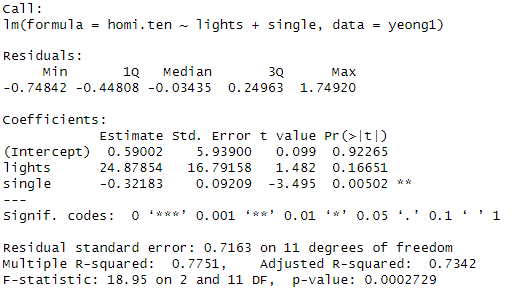


[그림2-4]전 처리 재 실시 후 절도 범죄 수 및 요인들

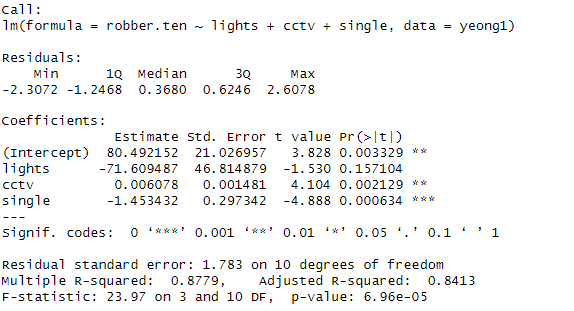


[그림2-5]전 처리 재 실시 후 폭력 범죄 수 및 요인들

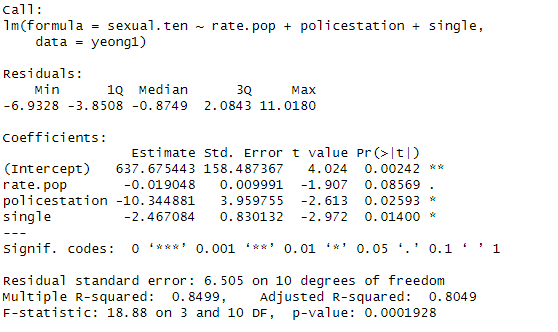
>영등포구



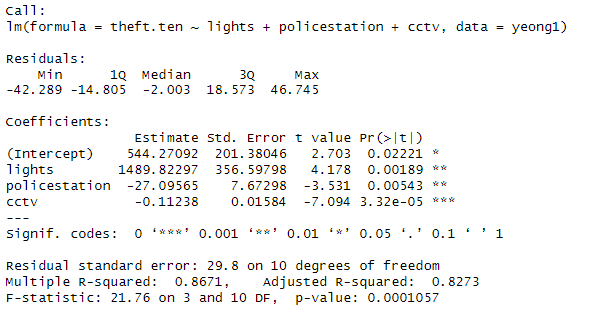
[그림3-1]전 처리 재 실시 후 살인 범죄 수 및 요인들



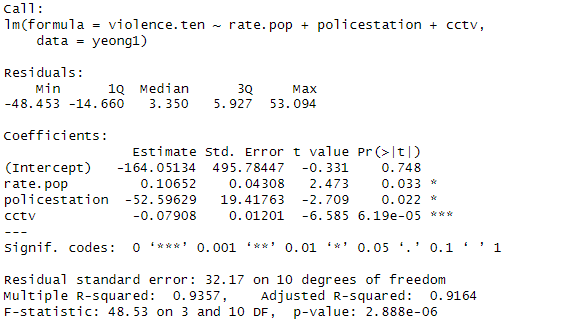
[그림3-2]전 처리 재 실시 후 강도 범죄 수 및 요인들



[그림3-3]전 처리 재 실시 후 성범죄 수 및 요인들



[그림3-4]전 처리 재 실시 후 절도 범죄 수 및 요인들



[그림3-5]전 처리 재 실시 후 폭력 범죄 수 및 요인들

분석 결과 해석:

각 구의 분석결과에 따르면

>살인: 정형화된 독립 변인만으로는 설명이 불가하다. 각 자치구마다 높은 적합성에도 불구하고 선별되는 변인들이 제 각 각인 점으로 미루어 보아 추가적인 요인들이 있을 것으로 보인다.

최인숙(2016)에 따르면 데이터를 기반으로 한 범죄정보 분석에는 5대 범죄(살인, 강도, 강간, 폭행, 절도)에 대해 범죄자의 성별, 나이, 교육정도, 생활정도, 직업별 빈도분석을 통한 범죄자 중심의 데이터를 이용한 유력 용의자의 상태를 산출하는 연구가 있고, 과거 범죄 발생 패턴을 근거로 공간에 대해 상대적 위험지수를 표현하여 위험도를 모델링하고, 시설물에 의한 상대적 위험도, 수목이나 하천에 의한위험도 등을 기본으로 위험도 확률 지도를 생성함으로써 시간의 흐름에 따라 달라질 수 있는 모델과 시뮬레이션 방법론을 제안했던 연구도 있다고 한다. 따라서 공간, 특히 시설적인 부분에서는 설명되지 않는 부분에 대한 충분한 합의가 가능하다.

>강도: 대체로 인구밀도, 보안등, 1인가구, 지역내 총 생산 과 관련이 있는 것으로 분석된다.

>성범죄: CCTV가 주된 요인이며 부가적으로 경찰서, 1인가구수가 유의미한 변인으로 분석된다.

>절도: 경찰 관서 수, 보안등 수, CCTV 수가 주된 요인이며 부가적으로 GRDP가 의미가 있다고 분석된다.

>폭력: CCTV가 주된 요인이며, 인구밀도, 주점 수, GRDP, 경찰 관서 수와 관계가 있는 것으로 분석된다.

경찰관서의 수는 높은 상관관계를 갖지 않다는 점에서 경찰관 수와 그에 따른 경찰 1명당 치안제공 가능한 시민의 수를 변인으로 선택해 추가적인 분석을 요한다.

그리고 GRDP의 경우 핵심 변인인 경우가 거의 없다는 점에서 보다 세부적인 항목으로 재설정 할 필요가 있는 것으로 보여 진다.

1. **활용 방안**

각 지역마다 높은 비중을 갖는 범죄유형이 있는데 이에 맞추어 억제력을 갖추기 위해 충분한 상관관계에 있는 요소의 설치 및 배치 과정에서 활용될 수 있으며, 반대로 현재 시점에서 지역 요인들의 정보에 기반하여 어떤 유형의 범죄에 충분한 억제력을 가지는지, 어떤 유형의 범죄에 취약한지 분석이 가능하다.

1. **참고 문헌**

최인숙(2016) 환경 특성자료에 기반한 범죄 발생 추이 분석, 디지털 포렌식 연구 제10권 제2호,66-80.