【붙임 2】

데이터 분석 최종결과보고서

Ⅰ. 참가자 정보

제 목	커널밀도분석을 활용한 보이스피싱 범죄 위험지역 예측모델 개발	
팀 명	통계민수	
성 명	김당찬	
연락처	휴대폰	010-2978-0937
	E-mail	dang11230@gmail.com

Kernel Density Estimation을 활용한 보이스피싱 범죄 위험지역 예측모형 개발

: 대전·세종 지역 신고 데이터를 바탕으로

TEAM 통계민수
김당찬, 강병국

목차

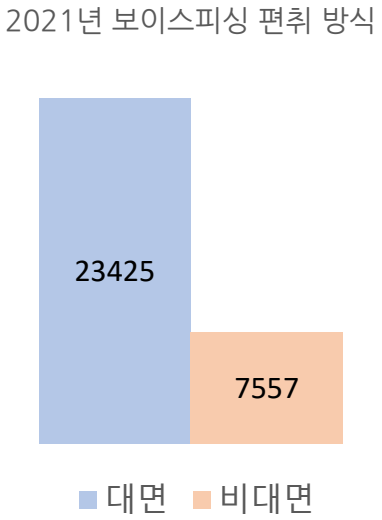
연구 배경

보이스피싱 범죄의 피해 규모는 대면 편취 위주로 증가하고 있으며, 적극적인 예방 대책이 요구됨

보이스피싱 범죄는 지속적인 수사 및 검거활동에도 불구,
범죄 수법이 다양해지고 피해액 역시 지속적으로 증가하는 상황*

특히, 송금 및 이체 등 비대면 편취 방식에 대해 정부가 예방책을 수립하면서 대면 편취 비율이 증가 추세*

대면 편취 방식은 주로 '고액 알바' 등을 명목으로 범죄 노출에 취약한 가출 청소년 등의 (현금)수거책 연루 문제 역시 발생



대면 편취 방식은 주로 피해자가 현금을 인출해 수거책에게 직접 혹은 간접적으로 전달하는 방식이므로
신속한 범죄 인지 및 선제적 신고 대응역량으로 **피해 경감** 가능
반면, 피해 후 사건 접수 및 수사로는 수거책 검거에 그쳐 **온전한 피해액 환수가 어려우므로**
적극적인 선제적 예방 대책이 요구됨

*' 보이스피싱 1건당 피해액 2500만원으로 늘어…코로나 이후 스미싱 피해 '급증', 조선일보 (2022)
https://biz.chosun.com/policy/policy_sub/2022/12/13/HKJIXTYSWFCI3L2NW2EFR6IHLA/?utm_source=naver&utm_medium=original&utm_campaign=biz

연구 배경

보이스피싱 범죄 예방을 위해 공간통계기법을 활용한 예측 모델을 생성할 수 있음

데이터 분석 및 모델링을 통해 보이스피싱 범죄를 분석하여 예방하기 위해서는 다음 방법들이 가능할 것

음성인식 / Text 분류 모델

보이스피싱 통화 음성 파일 혹은 스미싱 문자 텍스트를 바탕으로 피싱 유무를 분류하는 모델*

Spatial Analysis

보이스피싱 범죄의 피해 양상은 대면 편취 위주로 변화하고 있으므로, **공간통계기법**을 이용하여 구체적인 범행장소 예측

Fraud Detection

금융권 등에서 사용하는 사기거래 탐지(Anomaly detection) 모델을 이용해 보이스 피싱 데이터를 FDS에 융합하려는 시도가 있음**

본 연구에서는 신고일시 및 장소로 구성된 데이터를 기반으로
구체적인 보이스피싱 범죄 피해발생지역을 예측하는 모델 개발

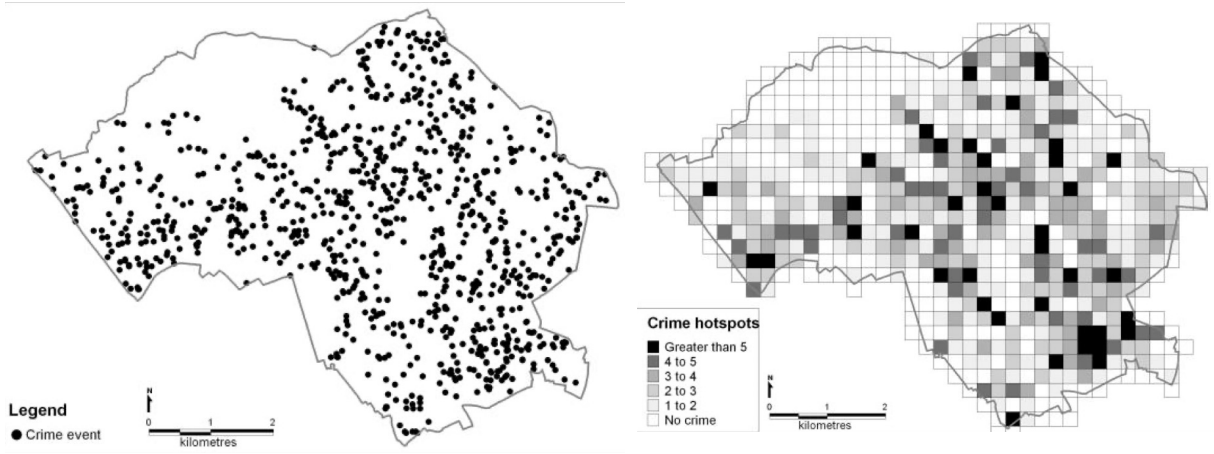
다만, 신고접수좌표가 현금 편취 장소와 일치하지 않을 수 있다는 문제가 존재
따라서, 신고 좌표와 사건 좌표가 (일정 오차 이내로) **일치한다는 가정** 전제로 분석

*ex. 전자금융사기 예측을 위한 문장 임베딩 기반 기계학습 적용에 관한 연구, 김정욱, (2020), ** 빅데이터와 FDS를 활용한 보이스피싱 피해 예측 방법 연구, 이승용 외, (2020)

분석 방법

커널밀도분석과 회귀모델을 바탕으로 범죄발생 위험지역 예측모델을 개발함

Kernel Density Estimation



Point Pattern(Left)을 바탕으로 격자 수준의 Density(Right)를 추정하는 모형*

각종 범죄의 발생 예측을 위해 Kernel Density Estimation 기반의 Hotspot 분석을 진행한 연구 다수 존재

- 범죄 발생 장소(Spot)만 고려한 Spatial KDE*
- 범죄 발생 시간까지 고려한 Spatiotemporal KDE**

Regression

Ex. Linear Regression Model

$$\mathrm{G}_i(\text{density of } i\text{th grid}) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

KDE로 구한 각 격자 별 추정 확률밀도를
외부 설명변수(인구통계학적 특성, 상업시설 여부 등)를
기반으로 예측하는 회귀모델(Ex. Linear Model) 개발

공간특성(인접행렬)을 반영한 공간회귀분석 모형 및
모델 성능 비교를 통한
높은 예측 성능을 가진 머신러닝 모델 생성
: 회귀모형을 바탕으로 세종 지역에 대한
위험지역 예측 및 비교

*The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime, Spencer Chainey et al. (2008)
**A Spatio-Temporal Kernel Density Estimation Framework for Predictive Crime Hotspot Mapping and Evaluation, Yujie Hu et al. (2018)

데이터셋 구성

제공된 신고데이터와 외부데이터를 이용해 타겟 데이터셋을 구성함(preprocessing.ipynb)

데이터 전처리는 Python의 Geopandas* 모듈 활용하여 공간정보 처리 후
외부 데이터와 병합하여 geometry를 포함한 공간데이터프레임 생성

*모든 공간데이터의 좌표계(crs)는 EPSG:5181 을 기준으로 함

분석 대상 데이터

신고 데이터

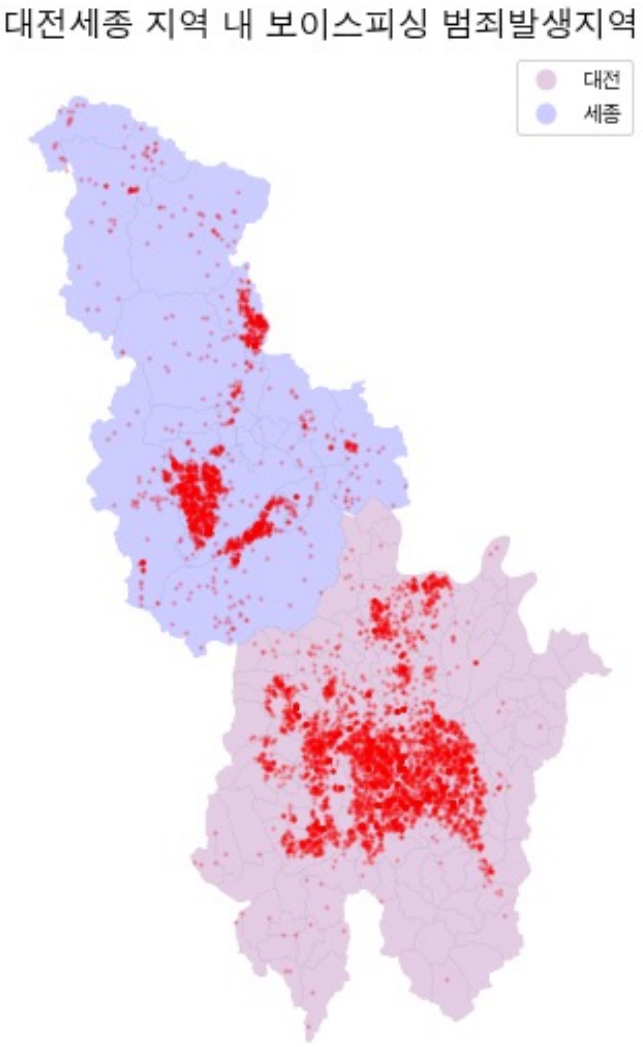
주어진 신고데이터 파일(NPA2020, KP2020, KP2021)
병합하여 아래 조건에 맞는 신고좌표 데이터 추출(오른쪽 plot)

- 대상 지역 : 대전광역시, 세종특별시
- 대상 일시 : 2021년도, 접수 날짜(일)만 추출
- 사건 코드 : 보이스피싱(`EVT_CL_CD == 215`)
- 동일사건 제외(`SME_EVT_YN != Y`)

격자 인구

통계지리정보시스템(SGIS) 에서 제공하는
각 지역별 격자(100M, 1KM 단위) 및
격자별 인구 데이터 이용

- 100M : 격자 내 총 인구, 남성인구, 여성인구 (3개 변수)
- 1KM : 격자 내 총인구, 연령대×성별 인구 (66개 변수)



데이터셋 구성

제공된 신고데이터와 외부데이터를 이용해 타겟 데이터셋을 구성함(preprocessing.ipynb)

분석 대상 데이터	상업시설	소상공인시장진흥공단 상가(상권)정보 공공데이터를 활용, 각 격자 내 포함된 상권종류별 상업시설 개수 파악(총 8종류) • 단, 소매시설 중 편의점은 ATM을 통해 현금인출이 가능하다는 점에서 별도로 파악
	은행	편의점 ATM을 제외하고 현금인출이 가능한 은행, 365코너(ATM) 등 • 동적 크롤러(Selenium 이용)를 개발하여 네이버 지도에서 각 시설의 주소 크롤링(Crawler_ATM.ipynb) • 지오코딩(Geocoding) 모듈인 geopy를 이용하여 각 시설의 좌표 추출
	버스정류장	국토교통부 전국 버스정류장 위치정보 공공데이터 활용, 각 격자 내 포함된 정류장 수 count

격자 크기(100m, 1km)에 맞는 두 종류의 데이터프레임 생성
대전 지역 데이터를 기반으로 모델링하여
세종 지역의 데이터로 예측 및 모델 성능 평가

Kernel Density Estimation

다음과 같은 방법으로 공간커널밀도분석 모형을 구성함

Spatial Kernel Density Estimation

$$\tilde{f}_h(\boldsymbol{y}|\boldsymbol{X}) = n^{-1}h^{-2}\sum_{i=1}^{n} K\left(\frac{\boldsymbol{y}-\boldsymbol{x}_i}{h}\right) q_h(\boldsymbol{y}|W)^{-1}; \qquad \boldsymbol{y} \in W,$$

Parameter $h(=h_s)$는 bandwidth로, 각 격자점에서 어느정도 거리까지의 사건을 포함할 지 설정한다

- 낮은 bandwidth : Undersmoothing(개별 point들만 반영)
- 높은 bandwidth : Oversmoothing
  (Unimodal distribution 형태가 됨)

커널 함수(Kernel) K는 2차원 공간 단위 KDE에서 주로 사용**하는 다음 Epanechnikov Kernel 함수 이용

$$K(u) = \frac{3}{4}(1-u^2) \text{ for } |u| \le 1.$$

Bandwidth Selection

R의 sparr패키지를 이용하여, 다음 가능도를 최대화하는 Bandwidth를 선택*
(Likelihood Cross Validation)

$$\mathrm{LIK}(h|\boldsymbol{X}) = n^{-1}\sum_{i=1}^{n} \log\left[\tilde{f}_h(\boldsymbol{x}_i|\boldsymbol{X}_{[-i]})\right].$$

이는 leave-one-out data(X_{-i})를 이용하므로 샘플 데이터로부터 직접 구할 수 있음

결과, Grid=100m 케이스에서
$h = 1179\ (m)$ 로
Spatial bandwidth 선택

* Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk with accompanying instruction in R, T.M. Davies et al. (2017)
** Hybrid Indexing for Parallel Analysis of Spatiotemporal Point Patterns, Alexander Hohl et al. (2016)

Kernel Density Estimation

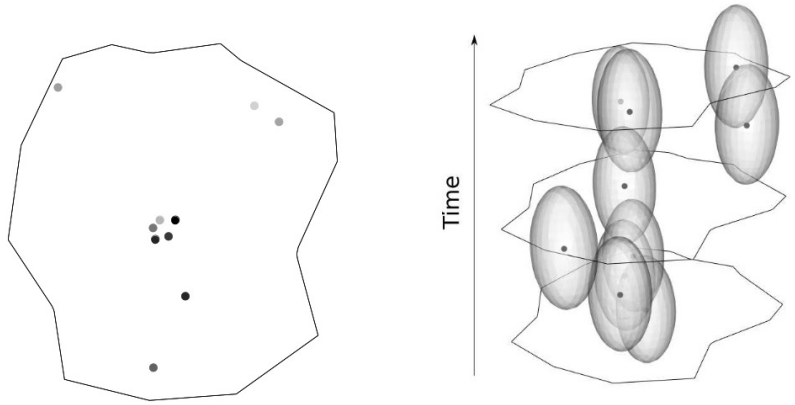
다음과 같은 방법으로 시공간커널밀도분석 방법을 구성함

Spatiotemporal Kernel Density Estimation

앞선 Spatial KDE에 추가적으로 데이터의 시간 차원(T)을 고려하는 모델로, 다음과 같이 추정 밀도가 주어짐

$$\breve{f}_{h,\lambda}(\boldsymbol{z}, s|\boldsymbol{\mathcal{X}}) = n^{-1}h^{-2}\lambda^{-1}\sum_{i=1}^{n} K\left(\frac{\boldsymbol{z}-\boldsymbol{x}_i}{h}\right) L\left(\frac{s-t_i}{\lambda}\right) q_h(\boldsymbol{z}|W)^{-1} w_\lambda(s|T)^{-1}$$

이는 각 신고 데이터의 공간적 분포 뿐 아니라 시간적 분포까지 포함하여, 한 격자 인근에서 데이터가 시간적으로 어떻게 밀집되어 있는지 고려(아래*)



Bandwidth Selection

선행연구**에서는 공간 범위 15km×20km 지역의 일일 데이터에 대해 100m의 공간격자와 $(2500\text{m}, 14\text{day})$의 bandwidth를 사용함

Spatial KDE에서와 마찬가지로, Likelihood-CV 기반 optimal bandwidth 선택 $(h_s = 299(m), h_t = 142(\text{days}))$

Oversmoothing bandwidth

추가적으로, 가능도를 계산하지 않고 경험법칙(rule of thumb)으로 oversmoothing bandwidth(h_{os})를 계산할 수 있음***

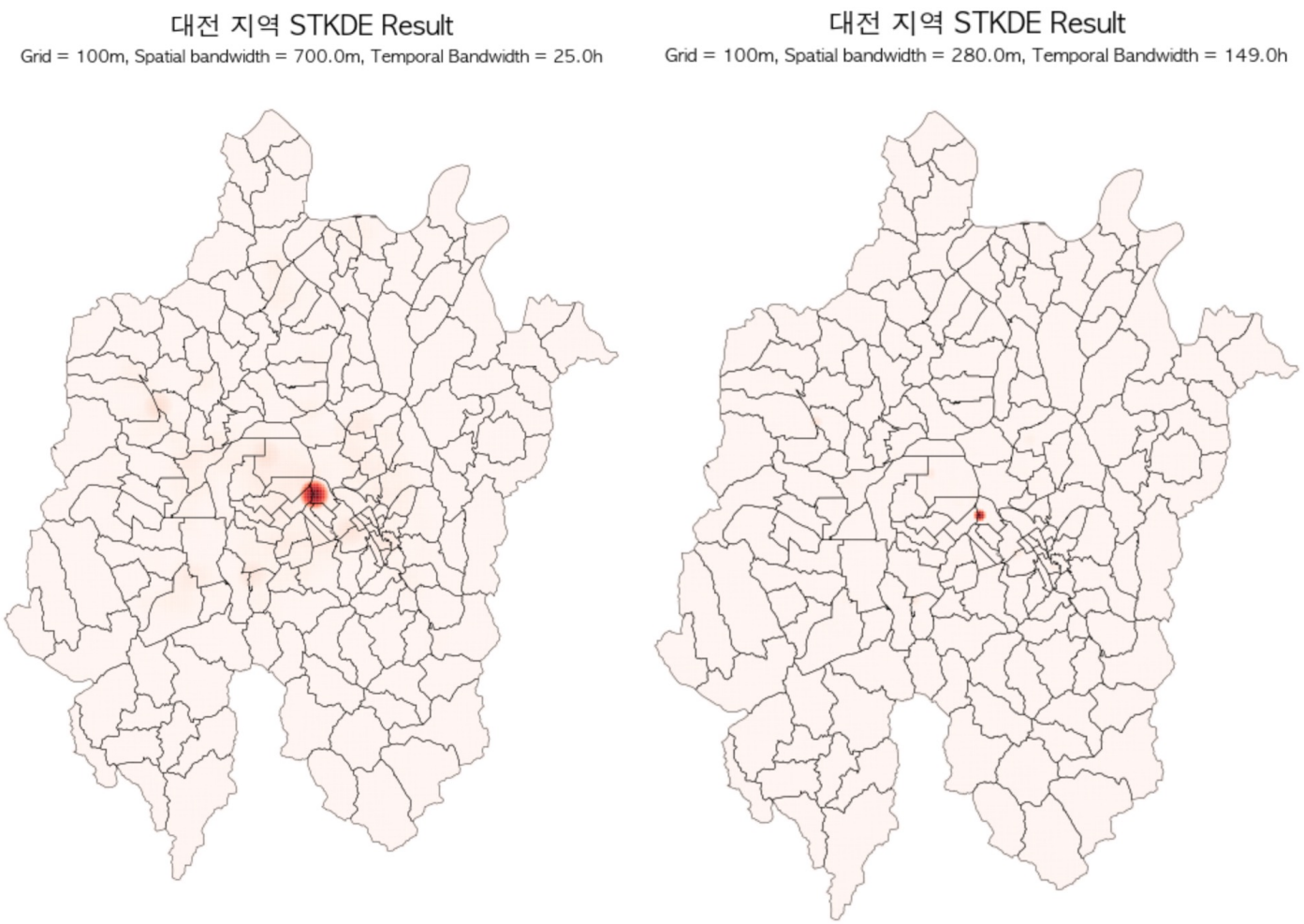
$$h_{\text{OS}(d)} = \hat{\sigma}\left\{\frac{(d+8)^{(d+6)/2}\pi^{d/2}R(K)}{16n\Gamma[(d+8)/2](d+2)}\right\}^{1/(d+4)}$$

*Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk with accompanying instruction in R, T.M. Davies et al. (2017)
** Hybrid Indexing for Parallel Analysis of Spatiotemporal Point Patterns, Hohl Alexander et al. (2016), *** The maximal smoothing principle in density estimation, G. R. Terrell. (1990)

Kernel Density Estimation - 결과

100m 격자 단위 시공간커널밀도분석으로 다음과 같은 결과를 얻음

Result (Grid=100m)



Grid = 100m의 Spatiotemporal KDE 결과
(붉은 부분이 $density > 0$인 격자를 나타냄)

Likelihood CV bandwidth(오른쪽)
: $h_s = 700, h_t = 25$
Oversmoothing bandwidth(왼쪽)
: $h_s = 280, h_t = 149$

Oversmoothing bandwidth을 이용해도 실제 사건분포를 반영하지 못하는데, 이는 보이스피싱 범죄가 특정 hotspot에만 시간적으로 Cluster 되어있음을 의미한다(뒷장 참고).

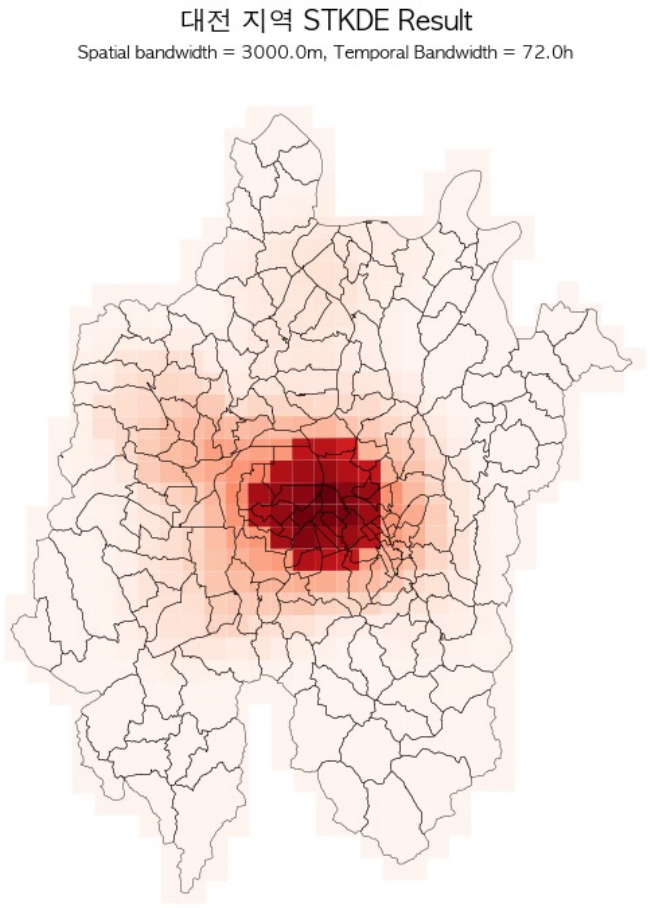
즉, 시간차원의 분석을 통해 **추가적인** hotspot을 반영하고자 하는 STKDE의 **목적*과 상충됨**

*A Spatio-Temporal Kernel Density Estimation Framework for Predictive Crime Hotspot Mapping and Evaluation, Yujie Hu et al. (2018)

Kernel Density Estimation - 결과

1Km 격자 단위 시공간커널밀도분석으로 다음과 같은 결과를 얻음

Result (Grid=1km)



Grid = 1km의 Spatiotemporal KDE 결과
(붉은 부분이 $density > 0$인 격자를 나타냄)

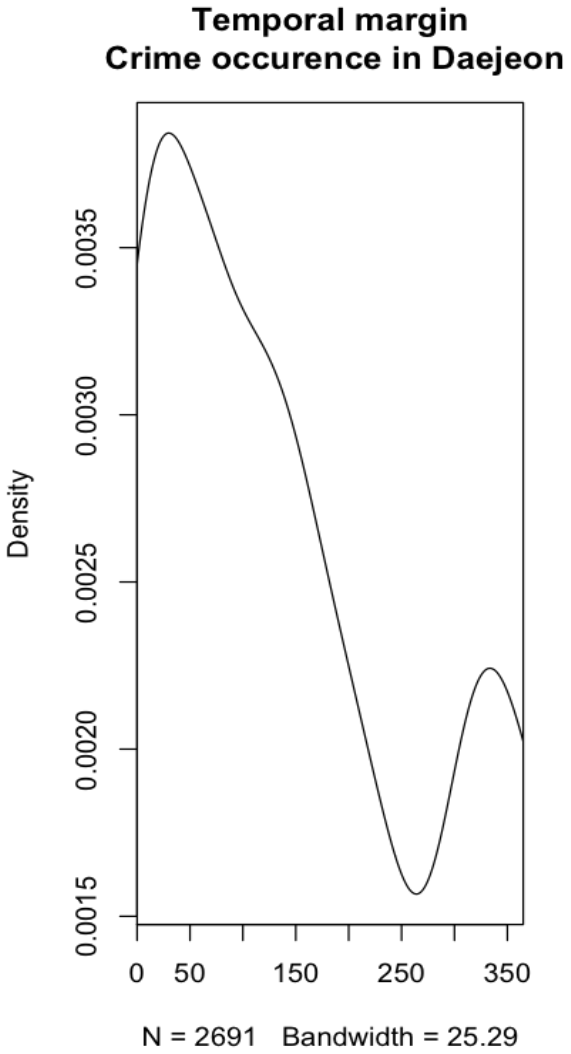
100M 격자 분석결과와 마찬가지로
중심 hotspot이 하나만 도출됨

다만, 격자 크기가 커짐에 따라
KDE 결과의 해상력(resolution)이 낮아져
추정 확률밀도가 continuous 하지 못함

STKDE Temporal Margin distribution*

: 각 일자별로(0-365) 시간차원의 분포를 확인

$T = 0 \sim 150$ 구간의 marginal distribution이
집중되어 있는데, 이로 인해 STKDE 결과의
hotspot이 이 시점에서의 사건들만 Capture 할
가능성이 높음

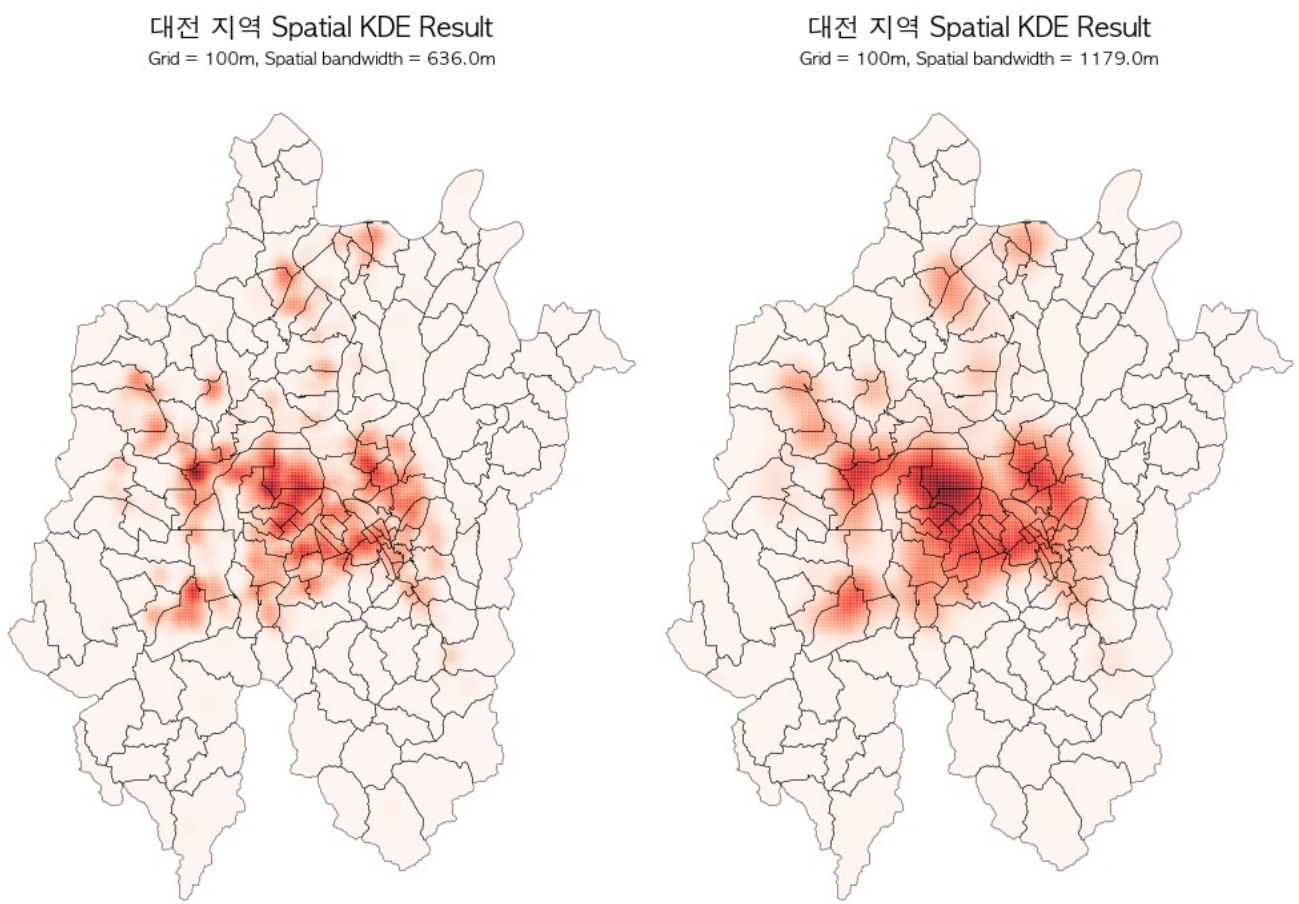


* R의 Sparr 패키지를 이용해 Plotting함

Kernel Density Estimation - 결과

100m 격자 단위 공간커널밀도분석으로 다음과 같은 결과를 얻음

Result (Grid=100m)



Grid = 100m의 Spatial KDE 결과
(Plot에서 붉은 부분이 $density > 0$인 격자)

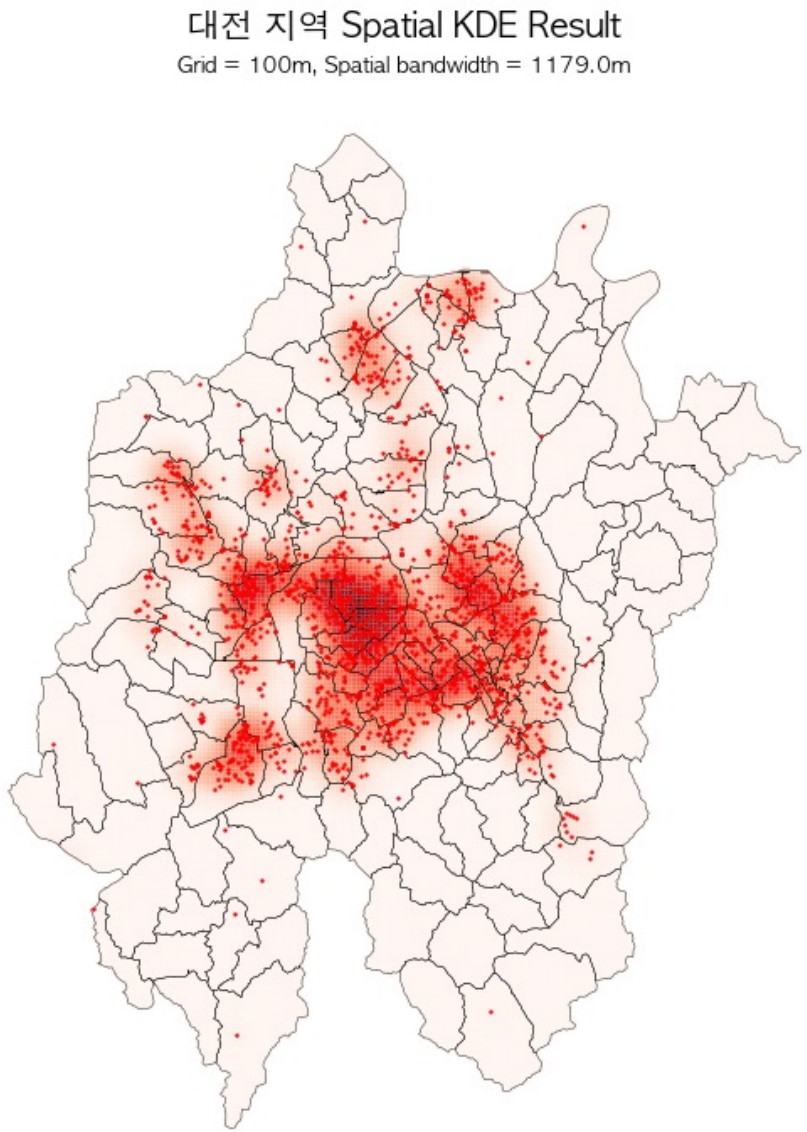
Likelihood CV bandwidth : $h_s = 636$
Oversmoothing bandwidth : $h_{OS} = 1179$
(STKDE와 같은 방법으로 h_{OS} 계산가능)

- Likelihood CV bandwidth 사용한 경우 hotspot이 전체적인 분포를 나타내기 보다는 개별 hotspot들의 집합으로 나타남
- Oversmoothing bandwidth 사용한 경우:
  개별 Hotspot의 영역이 커지면서 중첩되는 부분 발생하여 지역의 전반적인 분포로 확장

Oversmoothing bandwidth 이용이 Density Estimation에 적합하다고 판단

Kernel Density Estimation - 결과

최종적으로 공간커널밀도분석방법을 다음과 같은 조건으로 채택함



최종 선택 결과

여러 격자 수준 및 bandwidth 수준의 분석 결과,
최종적으로 **100m 격자 수준*에서 Spatial KDE 분석 결과 채택**
(bandwidth $h_s = 1179\text{m}$)

*1km단위 격자데이터에 많은 인구변수를 포함하더라도,
다중공선성 문제로 전체 변수를 사용하지 못할 뿐 아니라
표본크기 N 자체가 작기 때문에 부적합하다고 판단

대전지역의 보이스피싱 범죄 발생(2D-random variable)의
확률분포 $f(x, y)$는 왼쪽 그림과 같이 추정되며,
확률분포가 undersmoothing하지 않고, 외곽지역의
multimodal한 분포 역시 잘 capture함을 볼 수 있다.

(plot의 빨간색 점은 실제 신고 접수 좌표를 나타냄)

Regression - Spatial Regression

외생변수에 공간시차를 적용한 공간회귀모형을 통해 범죄 밀도 예측 모형을 개발함

공간자기상관성/다중공선성 진단

Multicollinearity

다중공선성 문제 : Variance Inflation Factor 계산
격자별 인구변수는 총 인구수(`to_in_001`)로 통일함

(격자별 인구변수 모두 포함시 VIF>1000으로 큰 다중공선성 발생함)

Spatial Autocorrelation

Moran's I Statistic 계산 : 각 변수 분포의 공간상관성 진단(표)

$$I = \frac{N}{W} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

Variable	I	z-score	p-value
density	0.998068	225.250569	0.001
ATM	0.006614	1.778677	0.046
Bus	0.068310	15.592823	0.001
Conv_Store	0.119085	28.371243	0.001
생활서비스	0.424673	98.177114	0.001
소매	0.207225	58.468786	0.001
학문/교육	0.306272	69.516582	0.001
음식	0.539817	127.205407	0.001
관광/여가/오락	0.329660	79.689763	0.001
부동산	0.200115	45.005985	0.001
스포츠	0.183857	44.132306	0.001
숙박	0.348588	75.628954	0.001
to_in_001	0.531114	129.279179	0.001

분석 결과, Cluster 개념으로 접근하는 density(highly clustered**)를 제외하고도

인구분포(to_in_001) 및 상권정보(음식점) 변수가 **0.5 이상의 높은 Clustering** 보임

: 외생변수들에 공간시차(spatial lag)를 설정하는 공간시차모형으로 회귀모델 구성

Spatially Lagged Exogenous Model*

$$\ln(P_i) = \alpha + \beta X_i + \delta \sum_j w_{ij} X_i' + \epsilon_i$$

특정 설명변수(X_i')에 대한 **공간 시차**(lag)를 설정하여 모델 구성
이때, Weight Matrix(W)로는 K-Nearest-Neighborhood 인접행렬 사용($k = 1$)

본 모델에서는
높은 Moran's I value를 나타낸
변수들에 대해 공간시차를 적용(시차변수 생성)

새로운 변수
`w_음식, w_to_in_001` 생성

* Pysal의 spreg 모듈을 이용하여 모델링함., **Moran's I가 1에 가까울수록 Highly clustered, 0에 가까울수록 Randomly dispered함을 의미, density의 경우 KDE 방식으로 추정한 데이터이므로 높은 공간상관성 보임

Regression - Spatial Regression

외생변수에 공간시차를 적용한 공간회귀모형을 통해 범죄 밀도 예측 모형을 개발함

Model Summary

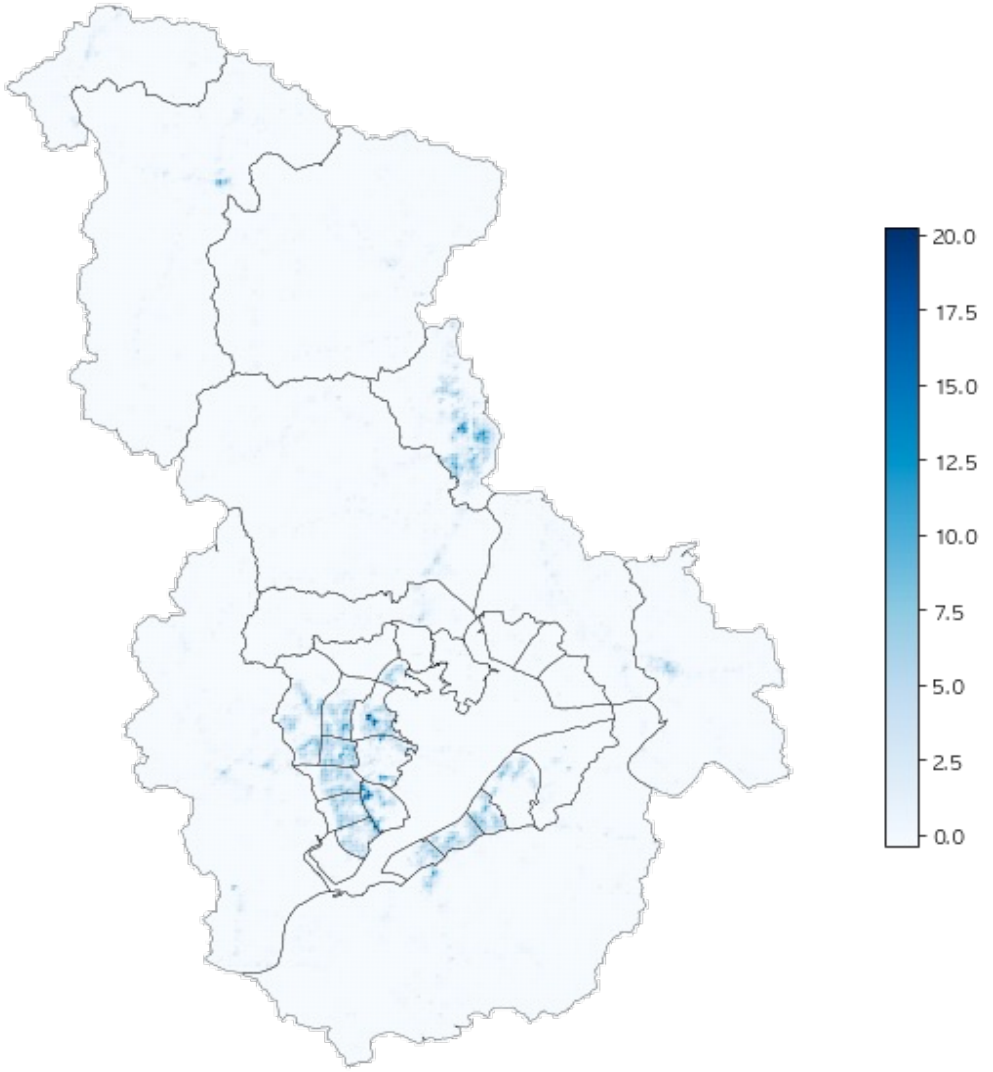
```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  : log_density                Number of Observations:       54910
Mean dependent var  :      0.0003                Number of Variables   :          15
S.D. dependent var  :      0.0006                Degrees of Freedom    :       54895
R-squared           :      0.4539
Adjusted R-squared  :      0.4538
Sum squared residual:       0.011                F-statistic           :   3259.6863
Sigma-square        :       0.000                Prob(F-statistic)     :           0
S.E. of regression  :       0.000                Log likelihood        :  346120.861
Sigma-square ML     :       0.000                Akaike info criterion : -692211.722
S.E of regression ML:      0.0004                Schwarz criterion     : -692078.020

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       0.0001855       0.0000020      92.2398134       0.0000000
                 ATM       0.0002801       0.0000198      14.1578755       0.0000000
                 Bus       0.0001299       0.0000067      19.4012750       0.0000000
          Conv_Store       0.0001147       0.0000122       9.3701197       0.0000000
            생활서비스       0.0000899       0.0000023      38.2689420       0.0000000
                 소매       0.0000038       0.0000007       5.5362697       0.0000000
              학문/교육      -0.0000190       0.0000030      -6.4263840       0.0000000
                 음식       0.0000193       0.0000014      14.1474526       0.0000000
            관광/여가/오락       0.0000059       0.0000079       0.7449783       0.4562880
               부동산       0.0000247       0.0000064       3.8751215       0.0001067
               스포츠       0.0000278       0.0000127       2.1955878       0.0281256
                 숙박       0.0002077       0.0000109      19.0220853       0.0000000
            to_in_001       0.0000013       0.0000000      56.5780892       0.0000000
                w_음식       0.0000500       0.0000010      48.6246217       0.0000000
          w_to_in_001       0.0000014       0.0000000      62.0804508       0.0000000
------------------------------------------------------------------------------------
```

Prediction

P-value가 유의한(<0.05) 변수들로 회귀식을 구성하여 예측대상인 세종지역 격자 데이터에 대해 회귀식을 적합

세종지역 예측 범죄 밀도(Normalized)



Regression - ML Modeling

Gradient Boosting Machine을 활용해 범죄 밀도 예측 모형을 개발함

ML Modeling 성능 지표 비교 결과

Model	MSE	RMSE	R2	RMSLE
gbr	**0.4745**	**0.6885**	**0.5262**	**0.2710**
lightgbm	0.4779	0.6910	0.5228	0.2711
catboost	0.4853	0.6963	0.5155	0.2726
rf	0.4985	0.7057	0.5023	0.2770
xgboost	0.5018	0.7081	0.4989	0.2761
knn	0.5155	0.7177	0.4853	0.2829
lr	0.5190	0.7201	0.4818	0.2807
ridge	0.5190	0.7201	0.4818	0.2807
lar	0.5190	0.7201	0.4818	0.2807
br	0.5190	0.7201	0.4818	0.2807
et	0.5566	0.7457	0.4442	0.2897

머신러닝 모델들의 Cross Validation Score 비교 결과 (일부 생략)

Tree 기반의 Boosting Machine인 Gradient Boosting Machine(Regressor)이 가장 높은 성능 (최소 예측오차 및 최대 R-squared)을 보임

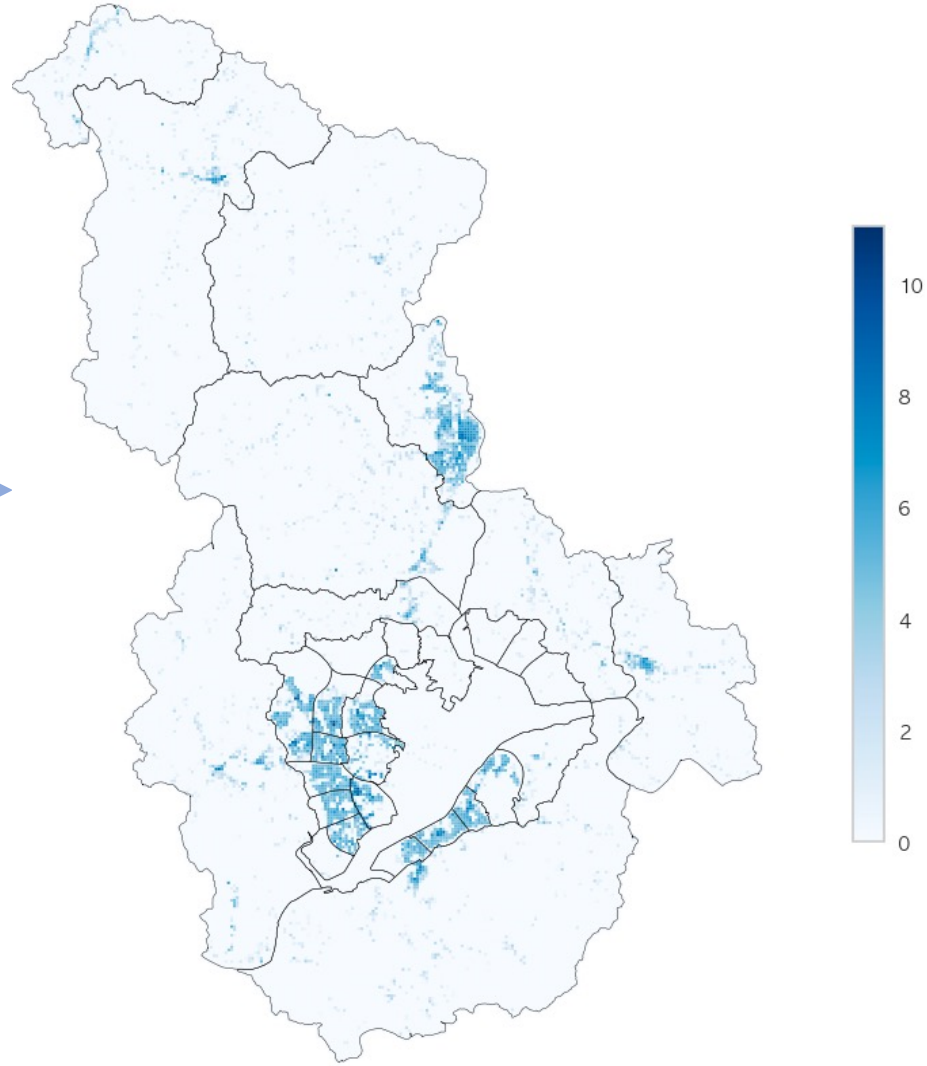
Bayesian Optimization*을 이용해 GBR model의 Hyperparameter Tuning
(lr=0.024, max_depth=3…)

* Scikit-Optimize 패키지 이용

Prediction

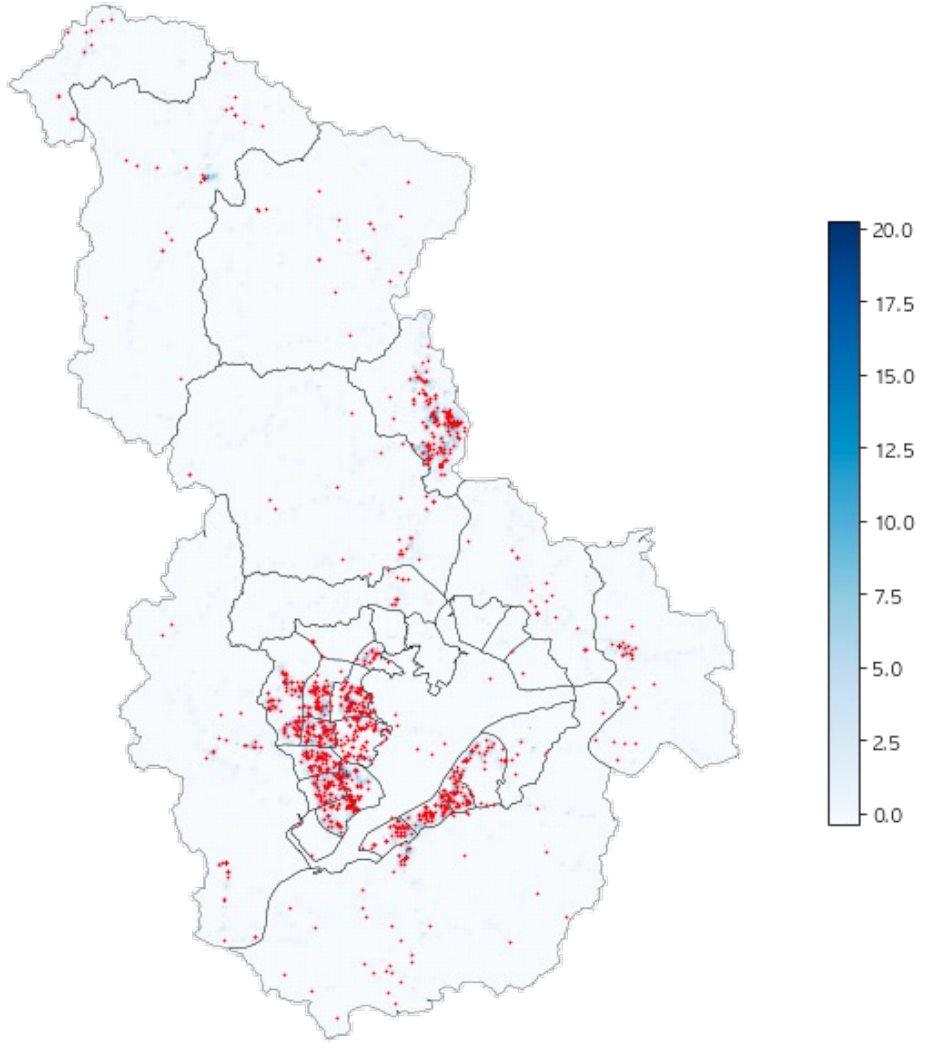
Tuning이 완료된 Gradient Boosting Machine에 예측대상인 세종시 격자 데이터를 대입

세종지역 예측 범죄 밀도(Normalized)

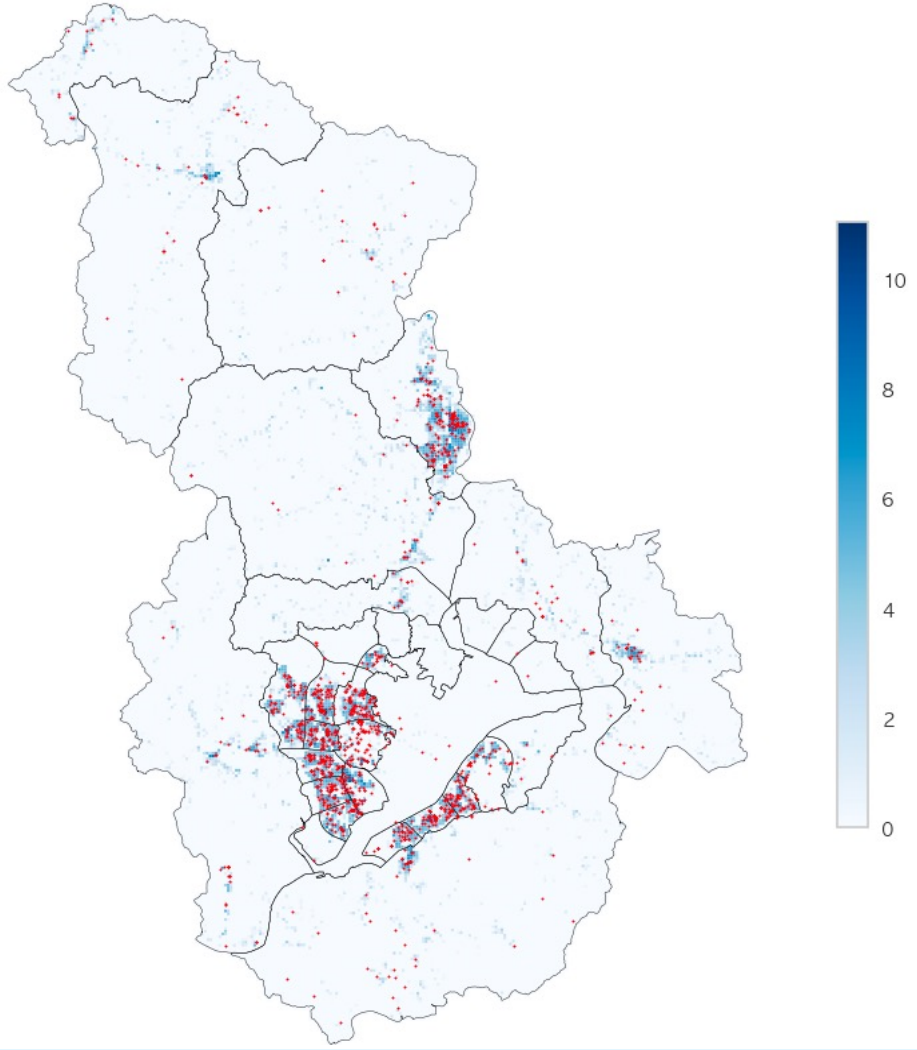


세종 지역의 실제 범죄 좌표 데이터와 비교 (spatial regression.ipynb, caret.ipynb)

Spatial Regression



Gradient Boosting Machine

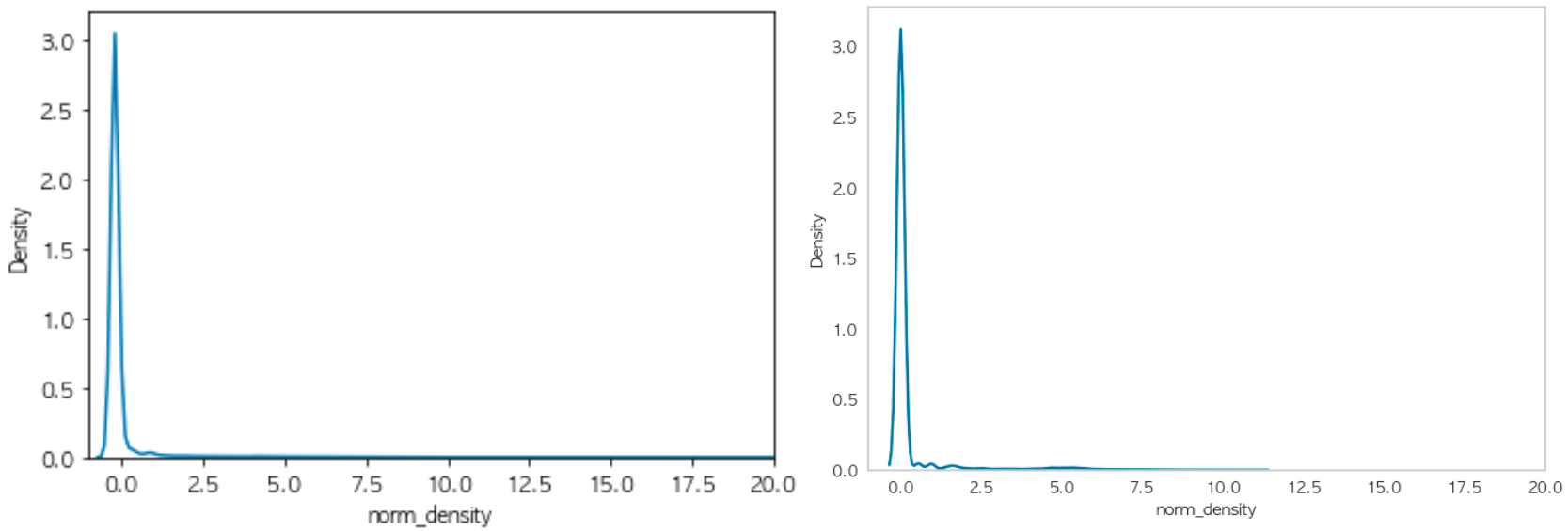


GBM에 비해 Spatial Regression 모델의 결과가 Outlying*한 격자에 의해 underfitting하는 것처럼 보이나*,
동일한 Recall** 값 대비 **spatial regression의 성능이 더 우수**함***을 확인할 수 있음

* 18페이지 predicted density distribution 참고, ** 18 페이지 Recall 참고, *** 19 페이지 참고

Spatial Regression Model vs. Gradient Boosting Regressor (spatial regression.ipynb, caret.ipynb)

Predicted Density Distribution



Spatial Regression으로 추정된 격자 밀도(normalized)의 분포(왼쪽)가
GBM으로 추정된 밀도분포에 비해 더 긴 tail을 가짐(Max = 20)
: 이로 인해 Plot에서 **Underfitting**이 이루어지는 것처럼 보임

Recall

$$Recall = \frac{TP}{TP + FN}$$

재현율(Recall)이란 실제 값이 True인 데이터($TP + FN$) 중
모델이 True로 예측한 데이터(TP)의 비율을 나타내는 성능 지표이다.

분류모델에서 주로 사용되는 지표이나,
본 연구에서는 실제 범죄 사건 중
예측 범죄발생격자에 **포함되는 비율**로
Recall을 정의할 수 있다.

Ex. 90% 재현율 예측에 필요한 격자 수 계산*

세종 지역 전체 격자 = 47377
전체 범죄 수 = 1422

Spatial Regression

90% Recall에 필요한 범죄 수 = 1276
90% Recall에 필요한 격자 수 = 12220 (**25.8%**)

Gradient Boosting Machine

90% Recall에 필요한 범죄 수 = 1276
해당 격자 내 포함된 범죄 수 = 28635 (**60.4%**)

* 예측 데이터셋으로부터, Estimated Density가 높은 순서대로 위험지역을 차례로 넓혀 나가는 방식을 이용함.

Spatial Regression Model vs. Gradient Boosting Regressor (model comparison.ipynb)

표. 세종시 격자 당 범죄 수 (내림차순)

	격자번호	범죄 수
0	29763	56
1	15604	41
2	9080	36
3	31852	36
4	8647	28
...	...	...
832	12500	1
833	12619	1
834	12620	1
835	12622	1
836	47293	1

세종시에서 발생한 2021년도
보이스피싱 범죄는 **총 1418개**
이들은 세종시 격자 47377개 중
837개에 분포되어 있음(표)

: 각 예측모델(공간회귀모형, Gradient Boosting Machine)이 예측한 density의 **누적분포(Cumulative Distribution)**를 바탕으로 두 모델의 성능을 비교*할 수 있음

그림 1. 누적 범죄 비율 당 사용된 격자 비율

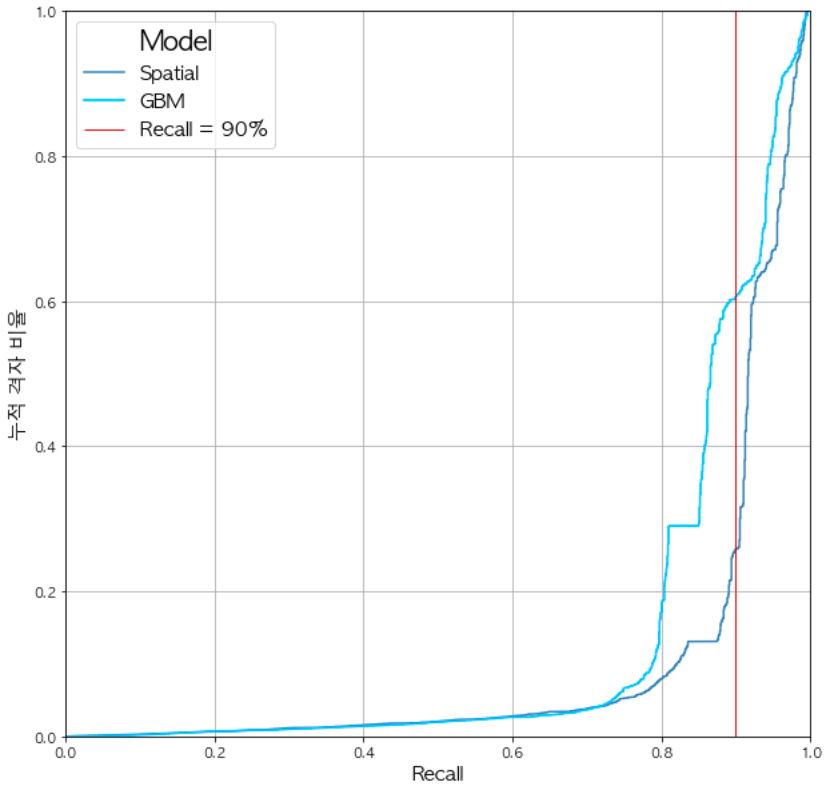
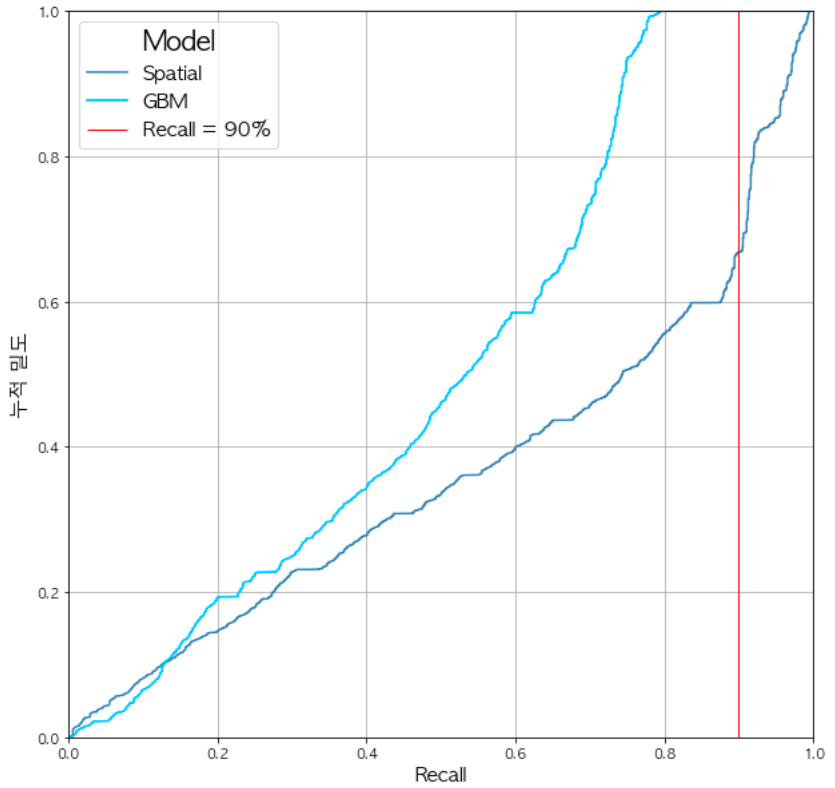


그림 2. 누적 범죄 비율 당 누적 밀도



(70% 이상의) 동일한 재현율($Recall = 0.9$, 붉은 실선)을 얻기 위해
Spatial Regression Model은 전체지역의 약 25%로 예측가능한 반면,
GBM Model은 전체지역의 60% 가량 및 누적 추정밀도 1.0 가량을 사용해야 예측가능

즉, 실제 모델 Fitting과정에서 R-squared 값은 GBM 모델이 더 높음에도 불구**,
재현율 측면에서 보면 **효율적인 예측 모델은 공간회귀모형**이라는 것을 알 수 있음

*각 격자 별 추정 밀도를 누적분포 형태로 비교하기 위해, 최소 추정밀도가 0이고 전체 지역 내 추정밀도 총합이 1이 되도록 Scaling함 **GBM=0.53, Spatial Regression=0.45

결론 및 기대효과

결론

제공 데이터셋과 외부데이터를 활용하여 대전 및 세종 지역의 보이스피싱 범죄위험지역을 **100m 격자 단위**로 분석한 결과
한 지역 내에서 보이스피싱 범죄가 발생하는 것을 **랜덤한 확률변수로** 가정하면

- **공간커널밀도분석**으로 oversmoothing bandwidth를 이용하여 지역 내 확률밀도함수를 추정할 수 있고
- **외생변수공간시차모형**(혹은 GBM 모델)을 활용하여 25.8%(60%)의 범죄 핫스팟으로 90%의 범죄 발생을 예측하는 모델을 생성할 수 있음

  Ex. 외생변수공간시차모형을 활용하여 얻은 세종특별자치시 내 보이스피싱 범죄 위험 상위 5개 행정구역은 다음과 같음(표)
  (n_grid : 각 행정구역과 겹치는 위험격자 개수)

기대효과

표. 세종특별자치시 내 보이스피싱 범죄 위험 상위 5개 행정구역

index	EMD_CD	EMD_NM	SGG_OID	GID	n_grid
29	36110250	조치원읍	36110	5126	563
23	36110340	금남면	36110	5120	347
28	36110360	연서면	36110	5125	337
24	36110350	장군면	36110	5121	333
13	36110112	고운동	36110	5110	23

본 연구의 방법과 같이 **공간밀도분석 및 회귀모형**을 이용하여 임의의 지역에 대한 보이스피싱 범죄 위험여부를 예측할 수 있음

만일 **개별 경찰관서**의 관내 지리정보 및 외생변수 데이터가 주어진다면 앞선 모델을 활용하여 해당 관내 보이스피싱 범죄위험 핫스팟 예측가능
: 이를 바탕으로 본청, 지방청 단위가 아닌 **개별 경찰서 단위**에서도 해당 지역을 중심으로 순찰 등 예방활동 강화 정책을 마련할 수 있을 것

또한, TAKDE*와 같이 커널밀도분석을 **실시간 신고 데이터**에 적용한다면 **real-time** 보이스피싱 예측 시스템 및 예방정책을 구현할 수 있을 것임

* TAKDE: Temporal Adaptive Kernel Density Estimator for Real-Time Dynamic Density Estimation, Yinsong Wang et al. (2022)

참고자료

참고문헌

- 보이스피싱 1건당 피해액 2500만원으로 늘어…코로나 이후 스미싱 피해 '급증', 조선일보 (2022)
- 전자금융사기 예측을 위한 문장 임베딩 기반 기계학습 적용에 관한 연구, 김정욱, (2020)
- 빅데이터와 FDS를 활용한 보이스피싱 피해 예측 방법 연구, 이승용 외, (2020)
- The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime, Spencer Chainey et al. (2008)
- A Spatio-Temporal Kernel Density Estimation Framework for Predictive Crime Hotspot Mapping and Evaluation, Yujie Hu et al. (2018)
- Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk with accompanying instruction in R, T.M. Davies et al. (2017)
- Hybrid Indexing for Parallel Analysis of Spatiotemporal Point Patterns, Alexander Hohl et al. (2016)
- The maximal smoothing principle in density estimation, G. R. Terrell. (1990)
- Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions - a crime case study. Monsuru Adepeju et al. (2016)
- TAKDE: Temporal Adaptive Kernel Density Estimator for Real-Time Dynamic Density Estimation, Yinsong Wang et al. (2022)

데이터 출처

- 격자 및 격자인구통계 : https://sgis.kostat.go.kr/view/pss/dataProvdIntrcn
- 버스정류장 : https://www.data.go.kr/data/15067528/fileData.do
- 상권 : https://www.data.go.kr/data/15083033/fileData.do

Code(Github)

- https://github.com/ddangchani/LocalSecurity_competition