Crime Data Analysis using Statistical Learning

EAS 508: Statistical Learning and Data Mining-I

Harish Sondagar, Akaash Vontivillu,
Riddhi Vaghani, Sidharth Saholiya

INTRODUCTION

- Crime patterns evolve over space and time, still most analysis focus only on past incidents.
- Static maps and historical count offer limited support for anticipating future risk.
- With limited law enforcement resources, early identification of high-risk areas is critical.
- This project aims to predict where and when crime risk may rise using spatio-temporal data and statistical learning methods.

DATA

- Primary Dataset: NYPD Complaint Data (2022–2024)
- Scale: Over 2.6 million crime incidents
- Attributes: Time, location (latitude - longitude), offense type, borough, and precinct
- Supplementary Data:

1. Socio-economic indicators (population, unemployment)
2. 311 service requests
3. Google Trends signals

- Structure: Spatially indexed and temporally aggregated for weekly analysis

METHODS / APPROACH

- Cleaned and validated raw crime records to ensure spatial and temporal accuracy.
- Discretised New York City into uniform spatial grid cells for scalable analysis.
- Engineered spatio-temporal features including weekly crime counts, lag features, and trends.
- Applied Kernel Density Estimation (KDE) to model continuous crime intensity and spillover effects.
- Used DBSCAN and HDBSCAN to identify persistent and emerging crime hotspots.
- Trained a Random forest classifier with probability calibration to predict future hotspot risk.
- Extended analysis to multi crime co-location and offense level risk pattern.

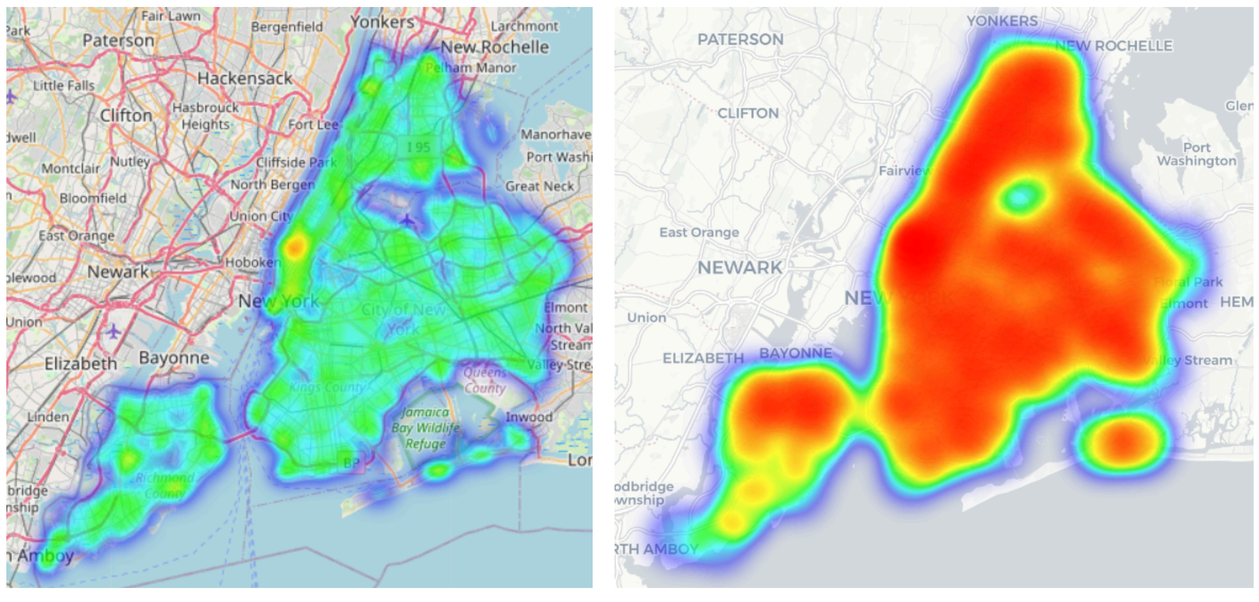


Fig: Ground truth vs Prediction

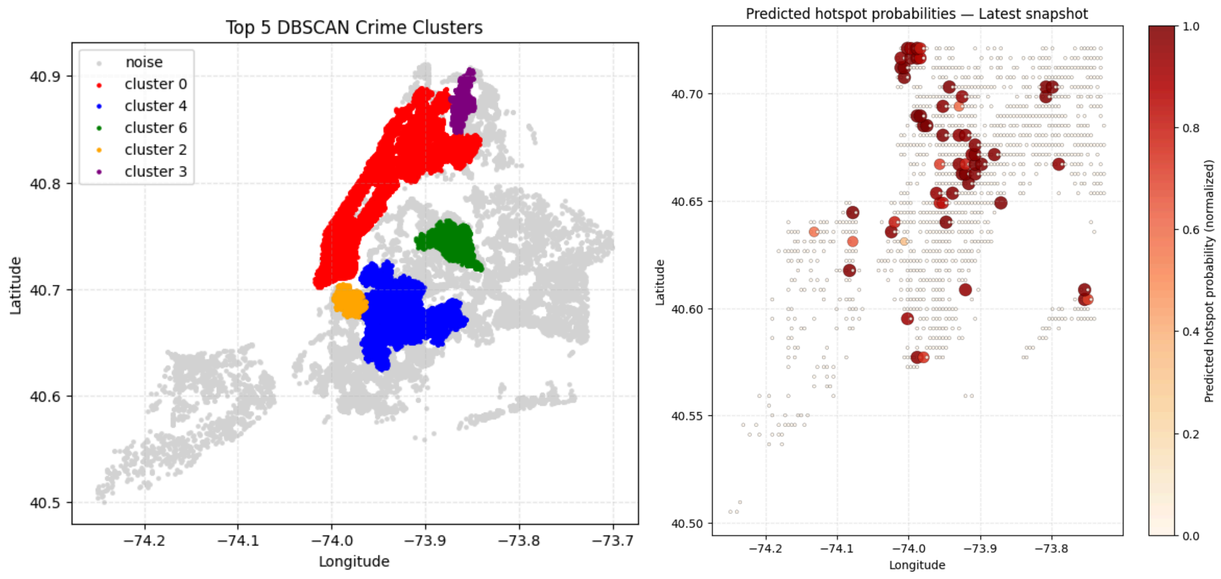


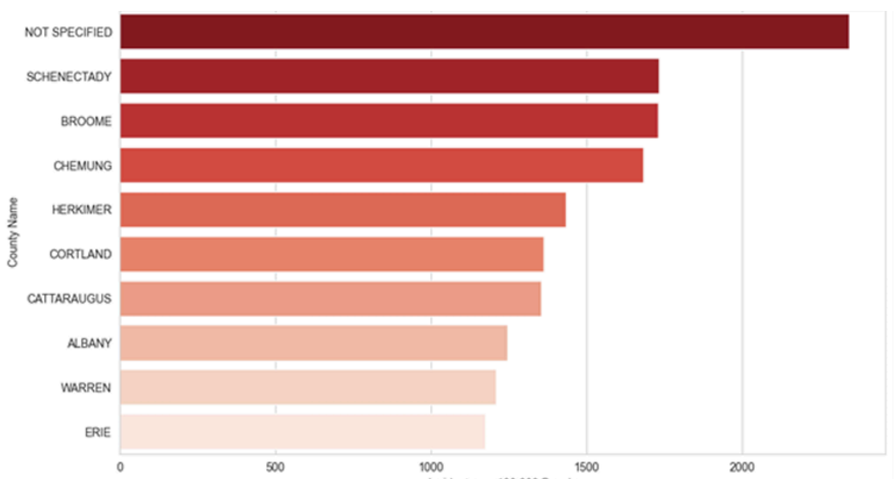
Fig: DBSCAN & Predicted hotspots

RESULTS & KEY FINDINGS

- KDE features improved spatial sensitivity, capturing near repeat and spillover crime effects.
- Clustering method identified constant hotspot zones that remained active across various past weeks.
- Multi-crime analysis revealed strong co-location between harassment, larceny and assault offenses.

Model	Performance
Precision@5%	99.5%
Brier Score	0.006
PAI@5%	16.3
Recall	98.6%

Table: Performance of model



Graph: Top 10 counties by crime rate

CONCLUSION & IMPACT

- Spatio-temporal modeling enabled early identification of emerging crime hotspots.
- Combining KDE, clustering, and machine learning improves prediction quality over static maps.
- Multi-crime hotspot analysis highlights areas requiring coordinated intervention strategies.
- Probability-based risk maps support more efficient and transparent resource allocation.
- The proposed framework is scalable and can be extended to other cities and real-time data sources.