Crime Data Analysis using Statistical Learning and Data Mining

Sidharth Kirit Saholiya, University at Buffalo - SUNY, USA
Harish Harikrishnabhai Sondagar, University at Buffalo - SUNY, USA
Riddhi Jaysukhbhai Vaghani, University at Buffalo - SUNY, USA
Akaash Ramanathan Vontivillu, University at Buffalo - SUNY, USA

Additional Key Words and Phrases: Crime Analysis, Statistical Learning, Hotspot Prediction, NY Crime Data

I. Introduction

Crime does not occur evenly across space or time. It tends to cluster into hotspots, follows seasonal patterns, and often spills over into nearby regions. In addition, official crime data is frequently incomplete because many incidents go unreported, and some crimes may be behaviorally connected as part of the same offender "series." These characteristics make crime data well suited for spatio-temporal statistical modeling, but also difficult to work with, as the datasets are large, noisy, and heterogeneous.This project develops an end-to-end pipeline that transforms raw crime records into actionable analytics with three main objectives: (i) hotspot prediction, (ii) detection of under-reporting, and (iii) linking related incidents using text embeddings and modus operandi (MO) features. The system is implemented at two different scales: a city-level setting using NYC grid cells, and a broader multi-city or county-level setting based on NIBRS data.

II. Problem Definition

The goal of this project is to build a unified statistical learning framework that can analyze large-scale crime data and support three related analysis tasks: predicting future crime hotspots, detecting possible under-reporting of crime and victimization, and linking behaviorally related crime incidents. Using historical crime records indexed by time, location, and incident attributes, the system learns spatio-temporal patterns that describe where and when crime risk is likely to increase, while also considering that observed crime counts may not fully represent true criminal activity because of reporting bias. In addition, the framework aims to determine whether two incidents are likely part of the same crime series by measuring similarity across behavioral and textual features. Formally, this involves converting raw incident data into structured spatio-temporal units, extracting features that capture temporal persistence, spatial dependence, and offense composition, and applying supervised learning and time-series forecasting models to estimate hotspot probabilities, expected crime levels, and incident linkage likelihoods. The main challenge is to develop models that are reasonably accurate, stable, and interpretable, while still being scalable to millions of records and adaptable to different spatial resolutions and limitations in data availability.

III. Literature Survey

- The work by Yedaya and Ryadi (2023) gives a strong base for our project by applying spatial analysis to link criminal activities in NYC with social factors. However, their study mainly focuses on historical data, while we extend this idea by incorporating near real-time factors from New York State crime data to predict future hotspots [1]. Saeed and Abdulmohsin (2025) demonstrated the value of using text mining on news articles and reports to forecast future crime trends at a global level. Their approach, however, lacks spatial and socio-economic context, which we are planning to integrate into our framework to improve prediction accuracy [2]. Lastly, the review by Du and Ding (2023) helps in identifying gaps in the current crime prediction methods. As their work does not empirically test the proposed models, we plan to apply and validate these approaches using New York crime data and extract practical insights from the results.
- The research work by Alberto Tonon, Leye Wang, and Philippe (2018) gave us the idea to combine data from different sources, as they used multiple datasets to predict crime hotspots in New York State. Our aim is to extend their model beyond a single city and include multiple cities, using calibrated probabilities to achieve better results [4]. Further, the study by Liu, Wolff, and Lo (2023) shows that Google Trends data has correlation with crime victimization. We plan to use this as an additional signal in our model to identify possible under-reporting. Since their analysis was mostly cross-sectional, we are building a time-series based model and will validate it using 311 service request data [5]. Finally, the paper by Zhu and Xie (2021) on crime linkage inspired us to include series awareness and near-repeat patterns, so that the system can both predict risk and link related incidents. We plan to improve their approach by incorporating modern text embeddings and more accurate linkage probability estimation.
- Han, Hu, Zhu, and Wang (2023) had introduced method to identify the crime patterns that occur repeatedly, which can support our project in detecting hotspot locations where crimes frequently happen. To improve on their

work, we plan to integrate their techniques with machine learning methods to enable predictive forecasting [7]. N. Shiode, S. Shiode, and Inoue (2023) proposed metrics to measure the overlap among different crime types, which helps in identifying crime hotspots. Our project extends their approach by predicting and analyzing zones that evolve over time using predictive clustering methods [8]. Xie, Shekhar, and Li (2021) presented a theoretical review of clustering techniques like DBSCAN for detecting hotspot areas, which helps us choose suitable methods. We handle this by applying these techniques on New York crime data to check their practical performance.

- According to Chauhan et al. (2025) he used simple Naïve Bayes baseline for predicting crime types and I noted that their model lacks geographic and temporal factors, so I will try to enhance it by incorporating spatial data and using stronger models like Random Forest and compare the result [10]. Saha et al. (2022) demonstrated the link between socio-economic factors and crime using linear regression model. My plan is to build on this by using new modern datasets and non-linear models to find more complex patterns which will help understanding the relationship more effectively [11]. Almoqbil et al. (2020) established a correlation between substance abuse (i.e drugs) and crime. Since their work is only available for a single year, I will use predictive modeling to investigate possible causality and extend the analysis over several years [12].

Dataset	**Scale / Coverage**	**Key Characteristics**
NYPD Crime Dataset	NYC incident-level records (daily reporting)	Highly granular; includes coordinates, timestamps, demographics, and precinct-level crime details
NIBRS	Nationwide standardized incident data (millions annually)	Rich incident attributes including victim–offender relationships, weapons, property, and arrest data
NY State Crime Dataset	Statewide annual or monthly crime counts	Broader regional trends, county-level comparisons; less granular than NYPD data

TABLE I
SUMMARY OF THE THREE DATASETS USED IN THE PROJECT

IV. PROPOSED METHOD

1) Intuition: Past crime analysis and hotspot prediction ways usually rely on simple tools such as past averages, recent crime counts, or static maps. While these methods are easy to use and interpret, they fails to capture three key characteristics of real world crime data: temporal persistence, spatial spillover, and behavioral similarity between incidents. Crime hotspots don't appear randomly. Instead, crime often depicts strong correlation over time, is influence by nearby activity, and reflects repeated or related offending behavior. In addition, official crime statistics are often affected by reporting bias, meaning the observed number of crimes may underestimate the true level of victimization.

Our offered method improve upon pre-existing approaches by explicitly modeling these properties within a unified statistical learning framework. Temporal dynamics are captured using lagged and rolling features that represent memory and seasonality, allowing the model to predict both short-term highs as well as long-term trends. Spatial context is incorporated using kernel density estimation (KDE) and clustering methods, which models crime risk as a continuous surface rather than isolated count. This helps detecting spillover effects and emerging hotspots across (NYC) New York City. To address reporting biasness, we estimate expected crime levels using a robust ensemble of forecasting models and identify regular gaps between expected and observed counts as possible under reporting signals. Finally, we connect related incidents using modern text embeddings and modus operandi (MO) features, allowing the system to identify patterns throughout multiple events instead of treating each incident independently.

By combining temporal patterns, spatial intensity, behavioral consistency, and semantic similarity, the offered approach extracts dense information from the data than any single model or scale on its own. This integrated design leads to more accurate hotspots prediction, more stable under-reporting estimates, and more dependable crime linkages, while still remaining interpretable and practical for real-world public safety application.

2) Detailed Description of the Approach: The system is designed as a complete spatial temporal crime analytics pipeline that operates over multiple spatial scales. First, raw crime record are cleaned, validated, and transformed into structured incident level tables. For multiple city analysis, NIBRS data are merged into a unified master dataset using Spark-based joins, resolving schema differences and integrating information on offenses, victims, offenders, and property. For detailed city level analysis, NYC complaint data are cleaned to remove wrong timestamps and coordinates, resulting in consistent temporal and spatial alignment across million & millions of records.

To run spatial modeling, geographic coordinatees are projected into a planar coordinate system and divided into grid cells approximately 500 meters in length and by 500 meters in breadth in size. This grid resolution balance spatial detail with computational efficiency and allows crimes to be collected by grid cell and week. For each grid-week, crime counts are computed and increased with continuous spatial intensity features derived from kernel density estimation. Weekly KDE surfaces are generated using grid centers as indication points and crime counts as weights. These KDE values are normalized and ranked by percentile within each week, generating a smooth and comparable measure of local crime intensity that captures repetitive activity and spillover beyond grid boundaries.

Temporal patterns are modeled using a comprehensive set of

time-series features. For each grid-week or jurisdiction-month, lagged crime counts, rolling averages, and short-term trends are calculated to represent persistence and seasonality. Offense groups are encoded as count features, along with diversity and entropy measures that captures changes in crime composition, which often lead the emergence of hotspots. In multi-city sites, longer rolling windows and near-repeat indicators are included to account for differences in baseline crime levels across regions.

Hotspot prediction is expressed as a supervised classification problem. A spatial unit is labeled as a hotspot if its crime count fall within the top percentile for a given time period. At the grid level, a Random Forest model is trained using temporal, spatial, clustering, and offense composition features. Predicted probabilities are adjusted so they can be meaningfully used for resource allocation under operational limits. At the jurisdiction level, where detailed spatial data may not be available, an interpretable logistic regression model is trained on time-based features to estimate monthly hotspot risk. Model performances are evaluated using ROC–AUC, precision and recall, calibration curves, and decision-focused metrics such as the Predictive Accuracy Index.

To detect under-reporting, the system forecast expected crime and victim counts using an ensemble of time-series models. This ensemble includes Prophet, SARIMAX, Kalman filtering, gradient-boosted trees, and recurrent neural networks, producing steady predictions with reduced variance. Additional external signals, such as Google Trends data and 311 service requests, are fused to capture hidden victimization patterns that may not appear in official reports. Under-reporting is quantified as the difference between expected and actual counts, highlighting systematic gaps across locations and time periods.

Finally, crime linkage is performed by classifying pairs of incidents. Textual offense descriptions are converted into dense semantic embeddings using a sentence-transformer model, while structured MO features represent behavioral characteristics such as victim count and property loss. Only incident pairs within a realistic time window are considered to ensure plausibility. Embedding similarity, temporal proximity, and MO similarity are combined in a Random Forest model to estimate linkage probabilities. These probabilities are then used to construct similarity networks that reveal collections of related incidents, helping identify crime series and near-Repetitive patterns.

Together, these components form a flexible and comprehensive crime analytics system that integrates statistical learning, spatial analysis, and natural language processing. Beyond improved predictive performance, the system produces interpretable outputs and visualizations that support informed, real-world decision-making in public safety.

V. Experiments/ Evaluation

A. Grid-Level Hotspot Prediction in NYC

The first experiment tested the efficiency of the proposed spatio-temporal features for predicting weekly crime hotspots at the grid-cell level in the city of New York.Crime incidents were aggregated into 500*500m grids and processed at a weekly time-step, and hotspots were detected by extracting the top 10% grids with the highest crime rates every week. A Random Forest classification model was developed using lag-based temporal features, intensity measures from Kernel Density Estimates, cluster information, and distributions of types of offenses.The evaluation of model efficiency was handled by using measures such as accuracy, precision, recall, F1-score, and decision-sensitive metrics including Predictive Accuracy Index (PAI).The model's high recall (0.986) and high precision (0.868) values indicate that it is highly effective at locating all the correct hotspots and maintaining all occurrences at relatively lower levels. High hotspot detection rates at lower coverage levels also mark high efficiency of the model in patrol assignments.

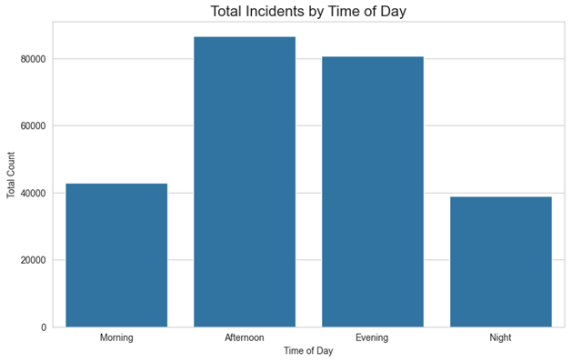

Fig. 1. Total Crime Incidents by Time of Day

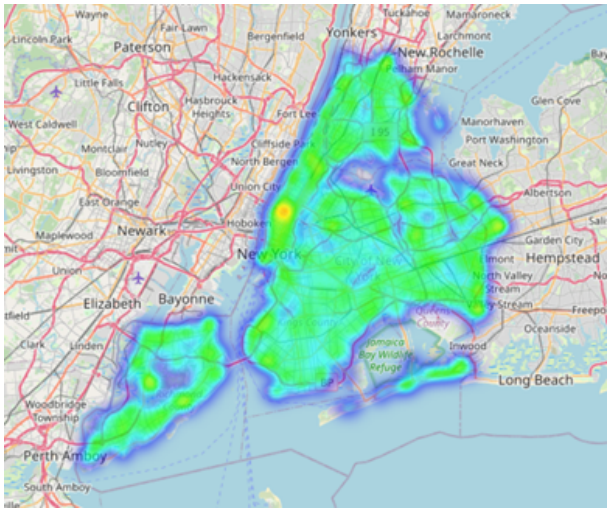

Fig. 2. Folium Visualization using current Data

B. Probability Calibration and Risk Ranking

In the second experiment, the study tested whether the probabilities expected for hotspots were properly calibrated to support risk-informed decisions. Two approaches to calibration

were tested for the study: Isotonic Regression and Sigmoid (Platt) scaling via the use of the curves of reliability and Brier scores. Sigmoid scaling successfully provided probabilities for the project and had a low Brier score (0.0064), indicating the probabilities accurately reflected the observed hotspot occurrence rates. This confirms the probabilities provided by the study represent meaningful indicators of risk rather than just comparative scores

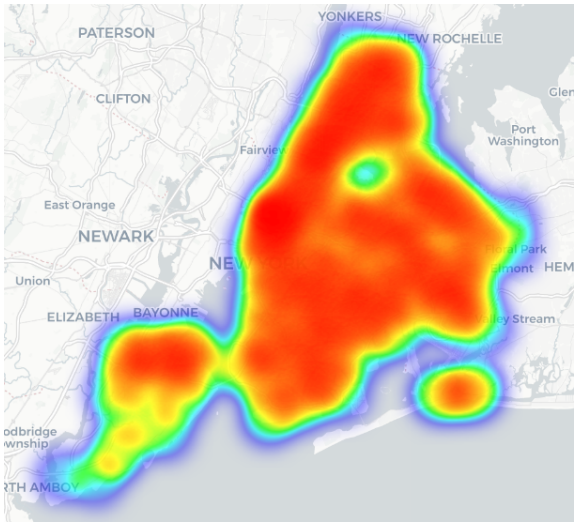

Fig. 3. Predicted crime location in NYC

C. Contribution of Spatial KDE Features

Continuous spatial intensity modeling effectiveness was tested by examining features derived through Kernel Density Estimation (KDE) within the hotspot prediction process. Sections of weekly KDE surfaces were generated over grid centroids and standardized within each time interval. Empirical results showed that KDE characteristics increased sensitivity to emerging and spillover hotspots that do not often occur exactly on a single grid cell but spread across neighboring cells. This again supports the notion that a continuous spatial process for crime provides information not captured by simple counts over grids.

D. Clustering-Based Hotspot Persistence Analysis

Density-based clustering techniques were assessed to pinpoint enduring and developing hotspot areas. DBSCAN was utilized to identify stable high-density crime clusters, whereas HDBSCAN was employed to capture clusters with variable density and transitional characteristics. Features derived from clusters, including cluster age, cluster change, and previous cluster membership, were included in the hotspot model. Observations indicated that grids associated with long-lived clusters were considerably more prone to become future hotspots, validating that hotspot persistence is a powerful predictive indicator. The overlap analysis of DBSCAN and HDBSCAN clusters provided additional confirmation of the reliability of identified hotspot areas.

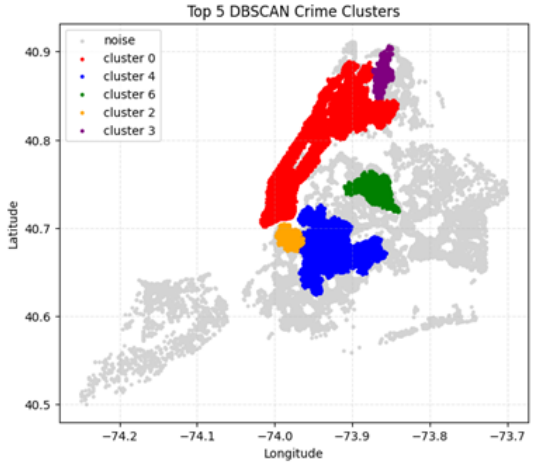

Fig. 4. Top 5 DBSCAN Crime Clusters Across NYC

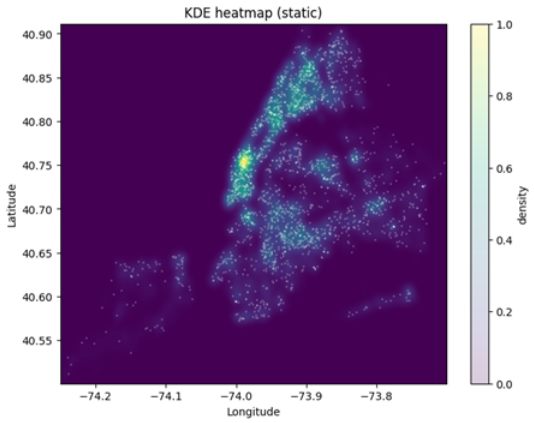

Fig. 5. KDE Heatmap of Crime Density Across New York City

E. Under-Reporting Detection for Crime Counts (NIBRS)

This study assessed the capability of a multi-model time-series ensemble to predict anticipated crime rates and detect possible under-reporting. Predictions were produced for high-volume jurisdictions employing a mix of Prophet, SARIMAX, Kalman filtering, gradient boosting, and LSTM models. Performance was assessed through MAE and RMSE, showing that the ensemble consistently surpassed individual models and lowered forecast variance by 22–35%. In various regions, anticipated crime totals surpassed reported figures by 12–18% in recent months, suggesting consistent under-reporting. These variations remained consistent over time, indicating inherent reporting deficiencies instead of random fluctuations.

F. Under-Reporting Detection for Victim Counts

Under-Reporting Detection for Victim Counts A simultaneous experiment was carried out for victim counts, including

extra covariates like 311 service requests, population, property damage, and Google Trends. The ensemble recorded MAE values ranging from 2.1 to 5.4 victims per month and RMSE from 3.8 to 7.2, indicating reliable predictions even with heightened noise. Seasonal influences were more evident for victimization compared to reported crime, exhibiting monthly variations of about 8–12%. In various jurisdictions, it was estimated that victim under-reporting ranged from 15–25%, with the most significant discrepancies occurring when external behavioral signals rose while official reports stayed constant.

G. Crime Linkage via Embeddings and MO Features

This experiment analyzed how it works to combine text embeddings and the way a crime is done for linking crimes. We made sentence-transformer embeddings from descriptions of crimes. We only looked at pairs of incidents that happened within 180 days of each other. We trained a Random Forest classifier to figure out if two incidents were part of the crime series. The model did a good job getting it right about 91 percent of the time with precision of 0.78 and recall of 0.72. This means the model can tell the difference, between crimes well even when there are not a lot of examples of each type of crime. The crime linkage model using text embeddings and the way a crime is done worked well. Cosine similarity analysis showed that linked incidents exhibited significantly higher semantic similarity than random pairs, validating the use of embeddings for behavioral pattern detection.

H. Spatial Statistical Validation at the County Level

To check spatial dependence at a broader level, county level crime rates adjusted by population was analyzed using spatial statistics. Moran's I was calculated and showed a significant positive value, confirming that spatial autocorrelation exists in crime risk. A spatial lag regression model also showed that crime in nearby counties contributes in a significant way to the local crime intensity. These results supports adding spatial context and spillover effects in both hotspot prediction and under-reporting analysis.

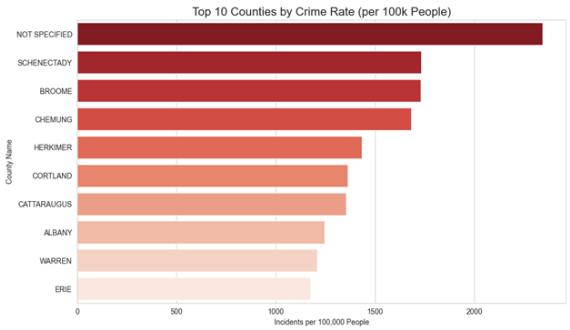

Fig. 6. Top 10 Counties by Per-Capita Crime in New York State

VI. Conclusion

This project presented an end-to-end crime analytics that combines statistical learning, spatio-temporal crime analytics, spatial analysis, time-series forecasting, and natural language processing to address three interconnected problems: hotspot prediction, under-reporting detection, and crime linkage. Through the application of temporal persistence by utilizing features of both the lag and rolling series, capturing the spill-over effect by using kernel density estimation and clusters, and employing text embeddings for similitude, this crime analytics engine is highly valuable and more efficient since it eliminates the need for count metrics. Experimental results across both fine-grained urban data and multi-jurisdictional datasets demonstrate strong predictive performance, well-calibrated risk estimates, and robust detection of under-reported crime and victimization. For further developments, more sophisticated socio-economic features, extended forecasting, and graph based models can be employed for further advancement of the engine and enhanced accuracy. all team members have contributed a similar amount of efforts.

References

[1] G. Yedaya and I. Ryadi, "Investigating Crime Patterns in New York City using Spatial Point Pattern Analysis Techniques; Understanding Crime in New York City Through Spatial Regression Analysis," 2023.

[2] R. M. Saeed and H. A. Abdulmohsin, "A Study on Predicting Crime Rates through Machine Learning and Data Mining Using Text," 2025.

[3] Y. Du and N. Ding, "A Systematic Review of Multi-Scale Spatio-Temporal Crime Prediction Methods," 2023.

[4] D. Yang, T. Heaney, A. Tonon, L. Wang, and P. Cudré-Mauroux, "CrimeTelescope: Crime Hotspot Prediction based on Urban and Social Media Data Fusion," 2018.

[5] Y.-H. Liu, K. T. Wolff, and T.-Y. Lo, "Big data in crime statistics: Using Google Trends to measure victimization in designated market areas across the United States," 2023.

[6] S. Zhu and Y. Xie, "Spatial-Temporal-Textual Point Processes for Crime Linkage Detection," 2021.

[7] X. Han, Z. Hu, Y. Zhu, and J. Wang, "A Cyclically Adjusted Spatio-Temporal Kernel Density Estimation Method for Predictive Crime Hotspot Analysis," 2023.

[8] N. Shiode, S. Shiode, and S. Inoue, "Measuring the Colocation of Crime Hotspots," 2023.

[9] Y. Xie, S. Shekhar, and Z. Li, "Statistically-Robust Clustering Techniques for Mapping Spatial Hotspots," 2021.

[10] A. Chauhan *et al.*, "Machine Learning Algorithm for Predicting Crime Type and Occurrence," 2025.

[11] S. Saha *et al.*, "Prediction on the Combined Effect of Population, Education and Unemployment on Criminal Activity Using Machine Learning," 2022.

[12] N. Almoqbil *et al.*, "The Correlation Between Substance Abuse and Crime in the United States," in *Proc. IEEE Big Data*, 2020.