Predictive Risk Modeling for Safety Interventions in Transportation Networks Using Spatial Crime History

Nazmus Sakib Pallab
Civil & Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
npallab2@illinois.edu

Jiarui Yu
Civil & Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
jiaruiy9@illinois.edu

Favour Jack
Civil & Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
frjack2@illinois.edu

Muhammad Fahad Ali
Civil & Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
mali19@illinois.edu

***Abstract*—Integrating personal safety into transportation and pedestrian planning requires systematic use of crime data. Information on crime location, time, and type can be analyzed to identify unsafe streets, intersections, and transit hubs, uncovering vulnerable areas in the urban network. Such insights enable engineers to propose design interventions such as reducing dead-end streets, improving pedestrian connectivity, and strategically relocating public transit drop-off points to enhance safety.**

In this study, raw crime record data will be transformed into actionable hotspot maps and predictive risk models to optimize the allocation of traffic police and patrol routes, ensuring coverage in the areas of highest need. Using advanced machine learning techniques, the study predicts crime types based on factors such as location, time of day, victim profile, and premises description. These results provide Civil Engineers and Urban Planners with evidence-based tools to prioritize infrastructure improvements and safety investments, while also identifying specific locations likely to evolve into future hotspots for proactive deployment of patrols, surveillance, and safety infrastructure.

Together, these predictive and spatial approaches are expected to enhance response efficiency and guide long-term city planning initiatives—from upgrading street lighting and redesigning public spaces to improving transit accessibility and targeting community resources—thereby strengthening the overall resilience and safety of urban infrastructure.

***Index Terms*—Transportation safety, Crime data analysis, Predictive risk modeling, Hotspot mapping, Machine learning, Urban infrastructure planning, Pedestrian safety**

I. Description of Dataset

The dataset used for this project is the ***"Crime Data from 2020 to Present"*** dataset for the City of Los Angeles, which is publicly available on DATA.GOV [1]. It is maintained and released by the Los Angeles Police Department (LAPD) as part of the city's open-data initiative, based on official crime reports filed by law enforcement officers.

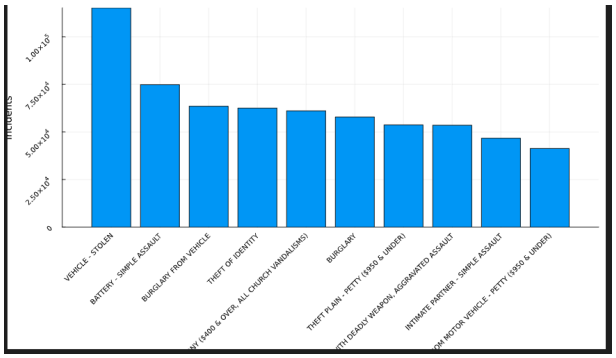
The dataset is provided in CSV format and contains over 1 million rows of crime incidents. The full dataset consists of 28 columns, while our project will focus on the following 12 key attributes:

- `DR_NO (Text)`: Division of Records Number: Official file number made up of a 2 digit year, area ID, and 5 digits.
- `Date Rptd (Floating and Timestamp)`: Date the incident was reported.
- `DATE OCC (Floating and Timestamp)`: Date the incident occurred.
- `TIME OCC (Text)`: Time the incident occurred.
- `AREA (Text)`: Area where the incident occurred.
- `AREA NAME (Text)`: ID of the area where the incident occurred.
- `Rpt Dist No (Text)`: A four-digit code that represents a sub-area within a Geographic Area.
- `Vict Age (Text)`: Age of the victim.
- `Vict Sex (Text)`: Sex of the victim.
- `Vict Descent (Text)`: Descent of the victim.
  - `LAT (Number)`: Latitude coordinate of the incident.
  - `LON (Number)`: Longitude coordinate of the incident.

This dataset provides both spatial (latitude/longitude, area, district) and socio-demographic (victim age, sex, descent) attributes, along with temporal information (date and time of crime occurrence), enabling spatial, temporal, and predictive risk modeling for transportation safety interventions.

II. Exploratory Data Analysis (EDA)

A. Crime Type Distribution



Group and rank by "Crm Cd Desc". Plot 1: bar chart of top 10 crime types. Insight: most frequent vs. least frequent crimes. (Favour Jack)

Check if image loads properly.

B. Temporal Patterns

C. Spatial Patterns

Another important aspect of our exploratory data analysis is the spatial distribution of crimes across Los Angeles. By mapping the latitude–longitude of each incident, we visualize hotspots and identify areas with high crime density.

We work with three spatial datasets: the set of crime points $C = \{C_i\}_{\{i=1\}}^{\{N_c\}}$, the set of street-lamp points $L = \{L_l\}_{\{l=1\}}^{\{N_l\}}$, and the city-boundary polygon B. We ensure longitudes and latitudes are numeric, finite, and within plausible ranges so that subsequent geometry remains meaningful.

We project all coordinates into a single metric CRS so distances are measured in meters:

$((x_i, y_i) = \Phi(\lambda_i, \varphi_i),\ (u_l, v_l) = \Phi(\lambda'_l, \varphi'_l))$

where:

- (λ_i, φ_i) and (λ'_l, φ'_l) are geographic (lon/lat, degrees) for crime C_i and lamp L_l,
- Φ denotes the projection to a local metric CRS,
- (x_i, y_i) and (u_l, v_l) are projected (planar) coordinates in meters.

A local projected CRS is used because Euclidean lengths in degrees are not physically meaningful, and a single metric CRS avoids unit mismatches.

We restrict the analysis to the jurisdiction by clipping points to the city polygon B, keeping only crimes and lamps whose projected coordinates fall inside B. This removes out-of-area points and reduces boundary artifacts that would otherwise inflate nearest-distance values.

We define planar Euclidean distance between a crime and a lamp:

$$d(C_i, L_l) = \sqrt{(x_i - u_l)^2 + (y_i - v_l)^2} \qquad (1)$$

For each crime, we keep its nearest-lamp distance: $d_i = m \underset{\{1lelleN_l\}}{\in} d(C_i, L_l), q \quad i = 1, ..., n$, where n is the number of crimes inside B. This yields a single, interpretable proximity value per incident. A spatial index keeps these queries fast.

From $\{d_i\}_{\{i=1\}}^{n}$ we build the empirical cumulative distribution function (ECDF): $F(r) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{d_i \le r\}, \quad r \ge 0$, so the coverage at an operational radius R is simply: $\text{Coverage}(R) = F(R)$.

We report robust summaries of proximity (median) and tail behavior (p90, p99) along with the maximum distance. Quantiles are preferred over the mean because the distribution is typically right-skewed and outlier-sensitive.

To guard against faulty records, we flag implausible distances above a conservative cap and, if needed, compute robust z-scores using the median and MAD to identify unusual d_i values.

Finally, to see how coverage accumulates with distance, we partition r into analyst-chosen bands (e.g., 0–50–100–250 m, …) and tabulate the share of crimes whose nearest-lamp distance falls into each band. This "coverage by band" view highlights where most gains occur (very small radii) and where diminishing returns set in as the radius grows.

d. Victim and Incident Attributes

4) Correlations and Relationships

5) Implications for Police Station Planning

Summarize the evidence:

Which areas have high and persistent crime density?

Which times need more coverage (e.g., night hours)?

Support with map + table of "Top 5 areas by crime density and trend".

III. Predictive Modeling

IV. Exploratory Data Analysis

A. Crime Type Patterns

B. Temporal Patterns of Crime

C. Spatial Distribution of Crime

D. Demographic Patterns of Crime Victims

For the whole dataset, we first analyzed the demographic distribution of crime victims based on age, sex, and descent. Overall, the victim population is 40.19% male, 35.68% female, and 24.13% unknown or missing.

V. Predictive Modeling

References

[1] Los Angeles Police Department / LAPD OpenData, "Crime Data from 2020 to Present." data.lacity.org / Data.gov, 2025.