# 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.

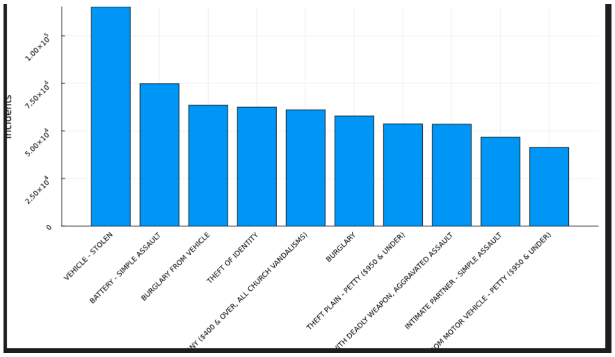
The dataset is provided in CSV format and contains over 1 million rows of crime incidents. The full dataset consists of 28 columns, while our project will focus on the following 12 key attributes:

- `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 examining the spatial distribution of crimes across Los Angeles. By mapping the latitude and longitude coordinates of each incident, we can visualize crime hotspots and identify areas with high crime density.

We have begun by ingesting and harmonizing three spatial datasets: the crime points $\mathcal{C} = \{C_i\}_{i=1}^{N_c}$, the street-lamp points $\mathcal{L} = \{L_\ell\}_{\ell=1}^{N_\ell}$, and the city-boundary polygon $B$. We have verified that longitudes and latitudes have been numeric, finite, and within plausible ranges so that subsequent geometry has remained meaningful.

We have projected all coordinates into a single metric CRS so that distance has been measured in meters:

$$(x_i, y_i) = \Phi(\lambda_i, \varphi_i), (u_\ell, v_\ell) = \Phi(\lambda'_\ell, \varphi'_\ell) \quad (1)$$

**where:**

- $(\lambda_i, \varphi_i)$ and $(\lambda'_\ell, \varphi'_\ell)$ have denoted the geographic (lon/lat, degrees) coordinates of crime $C_i$ and lamp $L_\ell$.

- $\Phi$ has denoted the projection from geographic to a local projected CRS with units in meters.

- $(x_i, y_i)$ and $(u_\ell, v_\ell)$ have denoted the projected (planar) coordinates in meters for $C_i$ and $L_\ell$.

We have adopted a local projected CRS because Euclidean lengths in degrees have not been physically meaningful, and a single metric CRS has prevented unit mismatches.

We have restricted the analysis to the jurisdiction by clipping points to the city polygon:

$$\mathcal{C}_B = \{C_i : (x_i, y_i) \in B\}, \mathcal{L}_B = \{L_\ell : (u_\ell, v_\ell) \in B\} \quad (2)$$

**where:**

- $B$ has denoted the study-area polygon in the projected CRS.
- $\mathcal{C}_B$ and $\mathcal{L}_B$ have denoted the subsets of crimes and lamps retained inside $B$.

We have carried out this step to remove out-of-area points and to reduce boundary artifacts that have otherwise inflated nearest-distance values.

We have defined planar Euclidean distance between a crime and a lamp:

$$d(C_i, L_\ell) = \| (x_i, y_i) - (u_\ell, v_\ell) \|_2 = \sqrt{(x_i - u_\ell)^2 + (y_i - v_\ell)^2} \quad (3)$$

**where:**

- $d(C_i, L_\ell)$ has been the straight-line distance (meters) between crime $C_i$ and lamp $L_\ell$.
- $(x_i, y_i)$ and $(u_\ell, v_\ell)$ have been their projected coordinates (meters).
- $\| \cdot \|_2$ has been the Euclidean norm.

We have chosen Euclidean distance because, under an appropriate local projection, it has approximated on-ground separation well while remaining computationally efficient.

For each crime, we have reduced to its nearest lamp distance: (

$$d_i = \min_{1 \le \ell \le N_\ell} d(C_i, L_\ell), i = 1, ..., n, n = | \mathcal{C}_B | \quad (4)$$

)

**where:**

- $d_i$ has been the nearest-lamp distance (meters) for crime $C_i$.
- $N_\ell$ has been the count of lamps inside $B$; $n$ has been the count of crimes inside $B$.

This reduction has produced a single, interpretable proximity value per incident and has yielded a one-dimensional distribution suitable for robust summaries. We have employed a spatial index so that queries have remained fast.

From $\{d_i\}_{i=1}^{n}$, we have constructed the empirical cumulative distribution function (ECDF):

$$F_{n(r)} = \left(\frac{1}{n}\right) \sum_{i=1}^{n} 1_{\{d_i \le r\}}, r \ge 0 \quad (5)$$

**where:**

- $F_{n(r)}$ has been the share of crimes whose nearest lamp has lain within radius $r$ meters.
- $1_{\{\{d_i \le r\}\}}$ has been the indicator (1 if the condition holds, 0 otherwise).
- $n$ has been the number of crimes analyzed.

We have then reported coverage at an operational radius $R$:

$$\text{Coverage}(R) = F_{n(R)} \quad (6)$$

**where:**

- $R$ has been the chosen policy/operational radius (meters); multiplying by 100 has given the percentage within $R$.

We have preferred the ECDF because it has been nonparametric, transparent, and directly tied to operational choices of $R$.

We have summarized typical proximity and tail behavior via quantiles and the maximum:

$$q_p = \inf\left\{x : F_{n(x)} \ge p\right\}, p \in \{0.50, 0.90, 0.99\}, d_{\max} = \max_i d_i \quad (7)$$

**where:**

- $q_p$ has been the $p$-quantile (meters) of the nearest-distance distribution.
- $q_{\{50\}}$ (median) has captured typical proximity; $q_{\{90\}}$ and $q_{\{99\}}$ have described the long tail.
- $d_{\max}$ has been the largest observed nearest-lamp distance (meters).

We have favored quantiles over means because the distribution has tended to be right-skewed; quantiles have resisted the influence of outliers.

We have flagged implausible distances with a conservative hard threshold:

$$n_{\{>T\}} = \sum_{i=1}^{n} 1_{\{d_i > T\}} \tag{8}$$

**where:**

- $T$ has been a geographic cap (e.g., $50\{,\}000$ m).
- $n_{\{>T\}}$ has been the count of distances exceeding $T$.

Optionally, we have standardized distances robustly using MAD:

$$\mathrm{MAD} = \mathrm{median}(\mid d_i - \mathrm{median}(d) \mid), z_i^{\{\mathrm{MAD}\}} \tag{9}$$

**where:**

- MAD has measured robust dispersion (meters).
- $\mathrm{median}(d)$ has been the sample median of the nearest-distance vector.
- $z_i^{\{\mathrm{MAD}\}}$ has been a robust z-score indicating how unusually large $d_i$ has been.

These checks have ensured that a handful of faulty records have not distorted interpretation.

We have quantified how coverage has accumulated across distance by partitioning the radius into bands. Given edges $0 = a_0 < a_1 < ... < a_K$, we have computed:

$$\Delta_k = F_{n(a_k)} - F_{n(a_{k-1})} = \left(\frac{1}{n}\right) \sum_{i=1}^{n} 1_{\{a_{k-1} < d_i \leq a_k\}}, \tag{10}$$

Note: $(k = 1, ..., K)$

**where:**

- $\{a_k\}_{k=0}^{K}$ have been analyst-chosen band edges (meters), e.g., $0, 50, 100, ....$
- $\Delta_k$ has been the fraction of crimes whose nearest-lamp distance has fallen inside band $k$.
- $F_{n(\cdot)}$ and $n$ have been as defined above.

We have also confirmed normalization over the covered range:

$$\sum_{k=1}^{K} \Delta_k = F_{n(a_K)} \leq 1 \tag{11}$$

**where:**

- $F_{n(a_K)}$ has been the cumulative share within the largest edge $a_K$.
- The inequality has held because ECDF values have not exceeded 1.

We have used bands because the vector $(\Delta_1, ..., \Delta_K)$ has acted like a discrete derivative of coverage, revealing where marginal gains have been concentrated and where diminishing returns have set in beyond small radii.

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.