Literature Survey

LI LINGYU, LIU YICHAO, WU JINGYAN, YANG QINGSHAN, YE FANGDA (Group 5)

Question Definition

- **Predictive Policing** is defined as *"the application of analytical techniques particularly quantitative techniques – to identify likely targets for police intervention and prevent crime or solve past crimes by making statistical predictions".*
- It can be categorized into 4 main classes:
  1. Predicting Places of Increased Crime Risk
  2. Predicting Potential Offenders
  3. Predicting Group/Population Crime Patterns
  4. Predicting Potential Victims

The main difference between them are:

Topic	Unit of Analysis	Key Driver	Ethical Risk
1. Places	Geography/Time	Environmental factors	Over-policing marginalized areas
2. Offenders	Individuals	Behavioral/demographic traits	Racial/profiling bias
3. Group Patterns	Networks/Populations	Social/economic interactions	Stereotyping demographic groups
4. Victims	Individuals/Groups	Vulnerability/exposure	Privacy violations

Column-to-Topic Mapping(from Chicago-Crime Dataset) & Possible Feature Engineering

Column	Relevant Topics	Feature Engineering Priorities	Key Concerns/Limitations
Date	1 (Places), 3 (Group Patterns)	Temporal features: `hour`, `day_of_week`, `month`, `holiday_flag`, `time_since_last_crime`	Seasonality effects (e.g., summer crime spikes).
Block	1 (Places)	Geospatial clustering (e.g., `hotspot_flag`), `distance_to_landmarks` (e.g., bars, transit hubs)	Partial redaction limits address precision.
IUCR	1 (Places), 3 (Group Patterns)	Encode crime type hierarchies (e.g., `violent_flag`, `property_flag`).	IUCR codes may change over time.
Primary Type	1 (Places), 3 (Group Patterns), 4 (Victims)	One-hot encoding for crime types (e.g., `THEFT`, `ASSAULT`).	Broad categories may obscure nuances (e.g., "theft" includes shoplifting and carjacking).
Location Description	1 (Places), 4 (Victims)	Categorical encoding (e.g., `residence`, `street`, `park`), `is_high_risk_location` flag.	Missing or vague entries (e.g., "other").
Arrest	2 (Offenders)	Temporal lag features (e.g., `prior_arrests_in_block`, `arrest_rate_per_district`).	Arrests $\neq$ crimes (biased policing may inflate counts).
Domestic	4 (Victims)	Flag for `repeat_domestic_incidents` (if address/block is repeated).	Underreporting of domestic incidents.

Column	Relevant Topics	Feature Engineering Priorities	Key Concerns/Limitations
Beat/District/Ward/Community Area	1 (Places), 3 (Group Patterns)	Aggregate crime counts by area (e.g., `crimes_per_1000_residents` using census data).	Community areas may proxy for race/income (ethical bias risk).
Latitude/Longitude	1 (Places)	Spatial grids (e.g., hexbin aggregates), `distance_to_police_stations`.	Coordinate accuracy varies (e.g., redacted blocks).
Year	3 (Group Patterns)	Long-term trends (e.g., `crimes_year_over_year`).	Limited utility if using finer-grained `Date`.

Summary of Existing Systems and Their Effectiveness

- **PredPol**, initially introduced in 2012, originated as a research project at the University of California. It has since evolved into the most widely implemented predictive policing algorithm in the United States. The system leverages three simple primary types of historical data: crime type, crime location, and crime time. By analysing these data, PredPol generates crime risk maps that indicates areas with high crime probability. The algorithm has proven effective in optimizing police resource allocation and reducing crime rates in certain areas. *(This paragraph may need to be simplified.)*

- **PredPol** uses methods called *place-based* or *hotspot* policing. Two fundamental models underlying PredPOl are ETAS (Epidemic-Type Aftershock Sequence) and the reaction-diffusion model. HunchLab, ShotSpotter, Oracle, Microsoft and so on also developed predictive policing systems.

  - **reaction-diffusion model**: describes the spread of crime in a city as a chemical reaction. The model assumes that crime spreads from one location to another in a similar way to how chemicals diffuse in a liquid.
  - **ETAS model**: treat the dynamic occurence of crime as a continuous time, discrete space epidemic-type aftershock sequence point process. The model is based on the idea that crimes are not isolated events but are influenced by past crimes in the same area.
  - Effectiveness: (TODO)
  - Drawbacks: (TODO)
    - statistical bias upon population.

- **Risk Terrain Modelling(RTM)** incorporates enviromental factors and sociological data by ML methods.

  - Effectiveness: (TODO)
  - Drawbacks: (TODO)

- Others like *people-based* as well as *group-based* policing uses a co-offence network to make predictions and recommendations. One example is gun violence prediction.

- *More to be filled here.*

  I think we better focus on PredPol and its theory first as HunchLab/CAS are just similar to PredPol. HunchLab is another popular predictive policing system in US but more complicated. It aims to predict how likely a particular crime type is to occur at various locations across a time period. HunchLab is based on machine learning and requires non-crime data. Same as PredPol, HunchLab will first collect crime related event data including location and time. The system can also process complicated geographic data contains points, lines and polygons, temporal data sets like weather data and school schedule, census data like house vacancy status, and natural terrain data.

  Crime Anticipation System (CAS) is a similar system used in Netherlands, and it relies on historical crime data to generate a heatmap to identify short-term crime risks. CAS operates on the principle that crimes are not randomly distributed but are concentrated in certain geographic and temporal clusters. The system integrates geographic clustering methods with temporal crime trends, recognizing that recent crimes often predict future crimes in nearby areas within a short time window. Same as PredPol, CAS will also collect basic crime data including type, location and time. Besides, CAS also process geospatial data like neighborhood layouts and temporal data like holidays and festivals.

Mathematical Details of PredPol

PredPol uses a straightforward statistical approach assuming that crime event follows a statistical distribution with two parameters: space and time. It believes past crimes are indicative of future crimes trends in nearby areas. #### 1. reaction-diffusion model In a discrete time model, taken *broken window effect* into consideration, we have [

$$B_s(t+\delta t) = \left[B_s(t) + \frac{\eta \ell^2}{z} \Delta B_s(t) \right] (1 - \omega \delta t) + \theta E_s(t) \tag{1}$$

] where $B_s(t)$ is a dynamic value models the risk of a site being attacked at time t. ω sets a time scale over which repeat victimizations are most likely to occur, and θ is a multiplier of $E_s(t)$, which is the number of burglary events that occurred at site s since time t. $0 \leq \eta \leq 1$ measures the significance of neighbor effect. l is the size of grid, z is the number of sites $s^{'}$ neighoring s .Δ is the discrete Laplacian, whereby [

$$\Delta B_s(t) = \frac{1}{\ell^2} \left(\sum_{s' \sim s} B_{s'}(t) - z B_s(t) \right) \tag{2}$$

] From the discrete model, we form the difference quotient:

[

$$\frac{B_s(t+\delta t) - B(t)}{\delta t} \tag{3}$$

]

and take the limit as ℓ and δt approach 0 to arrive at the differential equation

[

$$\frac{\partial B}{\partial t} = \frac{\eta D}{z} \nabla^2 B - \omega B + \epsilon D \rho A. \tag{4}$$

]

Here we have denoted

[

$$D = \frac{\ell^2}{\delta t}, \quad \epsilon = \theta \delta t, \quad \rho(s,t) = \frac{n_s(t)}{\ell^2}, \tag{5}$$

]

and ρ is the density of criminal agents

[

$$\frac{\partial \rho}{\partial t} = \frac{D}{z} \nabla \cdot \left[\nabla \rho - \frac{2\rho}{A} \nabla A \right] - \rho A + \gamma, \tag{6}$$

]

where offenders exit the system at the rate ρA and are reintroduced at the constant rate $\gamma = \Gamma / \ell^2$.

The PDE for ρ is obtained by a difference quotient for $ n_s(t) $, using the equation

[

$$n_s(t+\delta t) = A_s \sum_{s' \sim s} \frac{n_{s'}(t)(1 - p_{s'}(t))}{T_{s'}(t)} + \Gamma \delta t. \tag{7}$$

]

where [

$$T_{s'} = \sum_{s'' \sim s'} A_{s''}(t) \tag{8}$$

] which simply means that any agents that are present at s after one time step must have either arrived from a neighboring site after having not committed a crime there, or have been generated at s at rate Γ.

2. Using EM algorithm in ETAS model ETAS model assumes a continuous time, discrete space(square grids) problem. It supposes that each event generates $N \sim Possion(\theta)$ direct offspring events. In this model, policing areas are discretized into square boxes. The probabilistic rate of events in box n at time t is defined to be

[

$$\lambda_n(t) = \mu_n + \sum_{t_n^i < t} \theta \omega e^{-\omega(t - t_n^i)}, \tag{9}$$

]

where t_n^i are the times of events in box n in the history of the process. The background rate μ is a (nonparametric histogram) estimate of a stationary Poisson process.

The expectation, or E-step, sets

[

$$p_n^{ij} = \frac{\theta \omega e^{-\omega(t_n^i - t_n^j)}}{\lambda_n(t_n^j)}, \quad p_n^j = \frac{\mu_n}{\lambda_n(t_n^j)}, \tag{10}$$

]

where $\theta \omega e^{-\omega t}$ is called **the triggering kernel** that models “near-repeat” or “contagion” effects in crime data.

The maximization, or M-step, sets

[

$$\omega = \frac{\sum_n \sum_{i<j} p_n^{ij}}{\sum_n \sum_{i<j} p_n^{ij}(t_n^j - t_n^i)}, \tag{11}$$

]

[

$$\theta = \frac{\sum_n \sum_{i<j} p_n^{ij}}{\sum_n \sum_j 1}, \tag{12}$$

]

[

$$\mu = \frac{\sum_n \sum_j p_n^j}{T}, \tag{13}$$

]

where T is the length of the time window of observation. A more detailed derivation can be find here.

Review of the Modelling Approaches

Feature Engineering Techniques

Evaluation Metrics and Their Appropriateness