STATS 501 Project Proposal

Nina Bryan, Rachael Gu, and Marissa Fluharty

October 10, 2025

1 Team Members

Nina Bryan (Team Lead; ninabry@umich.edu), Rachael Gu, and Marissa Fluharty are M.S. Data Science candidates. Each will contribute to exploratory data analysis and one modeling method (GLM, GAM, or Spatio-Temporal Analysis).

2 Introduction

2.1 Motivation

The motivation for our project is to understand how spatial, temporal, and contextual attributes influence crime patterns in Chicago. By analyzing crime count data aggregated across neighborhoods and time periods, we can identify high-risk areas and temporal trends that inform public safety strategies. Building predictive models for crime counts will help us understand which factors most strongly contribute to crime concentration in specific areas and times.

2.2 Research Questions

The main questions we hope to answer from this analysis are:

1. Which neighborhoods (community areas) in Chicago experience higher rates of crime, and how does this vary across different crime types?
2. How do crime counts change across time (years, months, and potentially time of day)?
3. What is the relationship between location characteristics (e.g., location description, district, beat) and crime frequency?
4. Can we identify spatio-temporal hotspots that persist or shift over time?

3 Dataset

The dataset we have chosen, *Crimes 2001 to Present*, describes all reported crimes within the city of Chicago from 2001 to present (updated daily, minus the most recent seven days). This is provided by the City of Chicago's data portal and is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In addition, the city also provides a dashboard with visualizations that will assist in preliminary exploration.

Records with missing geographic identifiers (District or Community Area) will be excluded from the analysis, as these represent a small proportion of the data and cannot be reliably assigned to spatial units.

4 Proposed Methods

4.1 Generalized Linear Model

A GLM model will be useful for modeling the response variable, crime counts, and for exploring various predictors such as crime type, location, and time. Since crime data often exhibit substantial variation due to differences in location and time, we will use a model that can handle overdispersion. A quasi-Poisson or negative binomial distribution with a log link allows us to examine how crime frequency varies across locations and crime types. An example of the GLM model is:

An example GLM specification:

$$\log\left(E[\text{CrimeCount}_{it}]\right) = \beta_0 + \beta_1 \text{CrimeType}_i + \beta_2 \text{District}_i + \beta_3 \text{Year}_t + \beta_4 \text{Month}_t$$

Overdispersion diagnostics: We will assess overdispersion using:

- Deviance and Pearson's χ^2 statistics, checking if residual deviance $\approx \text{df} \pm 2\sqrt{2 \times \text{df}}$
- Comparison of Poisson vs. Negative Binomial models
- **Note:** AIC comparisons between Poisson and Negative Binomial may not be directly comparable under overdispersion

If overdispersion is detected, we will use a Negative Binomial GLM or quasi-Poisson approach to obtain appropriate standard errors.

4.2 Smoothing/Semi-Parametric Model

This approach will employ a **generalized additive model (GAM)** using penalized splines to capture nonlinear relationships in crime counts. We will use the `mgcv` package in R with REML estimation for smoothing parameter selection.

The model specification:

$$\log(E[Y_{it}]) = \beta_0 + s_1(\text{longitude}_i, \text{latitude}_i) + s_2(\text{time}_t) + \beta_1 \text{CrimeType}_i + \beta_2 \text{LocationDesc}_i$$

where:

- Y_{it} denotes the crime count in area i at time t
- $s_1(\cdot)$ is a smooth spatial surface using penalized splines (e.g., thin-plate or tensor product splines)
- $s_2(\cdot)$ is a smooth function of time to capture temporal trends
- Parametric terms include categorical predictors like crime type and location description

Implementation details:

- Use Poisson or Negative Binomial family depending on overdispersion diagnostics
- Fit using `gam()` function from `mgcv` with `method="REML"`
- The connection between mixed effects models and penalized splines will be leveraged through REML estimation
- Model selection will use REML, GCV scores, AIC, and deviance explained

We will compare the GAM with the parametric GLM to assess whether the added flexibility of smooth terms significantly improves model fit and predictive performance.

4.3 Spatio-temporal Analysis

The third approach extends the GAM by incorporating **random effects** to account for spatial and temporal correlation structures. This mixed-effects framework (GAMM) will be implemented using `mgcv` or `gamm4`.

Model specification:

$$\log(E[Y_{it}]) = \beta_0 + s_1(\text{longitude}_i, \text{latitude}_i) + s_2(\text{time}_t) + \mathbf{Z}_{it}^\top \boldsymbol{\beta} + b_i + u_t$$

where:

- $b_i \sim \mathcal{N}(0, \sigma_b^2)$ captures unobserved spatial heterogeneity across neighborhoods (random spatial effect)
- $u_t \sim \mathcal{N}(0, \sigma_u^2)$ captures temporal correlation (random temporal effect)
- $s_1(\cdot)$ and $s_2(\cdot)$ are penalized smooth functions
- $\mathbf{Z}_{it}$ represents parametric predictors

Key features:

- Random effects account for correlation among observations within the same community area over time
- Penalized splines will be estimated using the mixed model representation, where smoothing parameters are related to variance components via REML
- This framework allows joint inference on both smooth surfaces and random effects
- We can assess whether spatial and temporal random effects significantly improve model fit beyond the smooth terms alone

4.3.1 Expected Outcomes

We expect:

- The GLM to reveal significant linear associations between crime counts and predictors like district, crime type, and time
- The GAM to uncover nonlinear patterns in space and time, such as crime hotspots and seasonal trends
- The GAMM to demonstrate that accounting for spatial and temporal correlation improves model fit and inference, revealing how crime concentrates and shifts across Chicago neighborhoods over time

Together, these methods will provide interpretable inference, robust prediction, and strong alignment with the GLM, mixed effects, and semiparametric modeling objectives.

Appendix

A Project Timeline

- **October 10, 2025 (Proposal Submission)**
  Submit a 1–2 page proposal outlining the project goals, data source(s) (Chicago crime), and planned methods (GLM, GAM, and GAMM). Finalize team roles and confirm data access.
- **October 11–24, 2025 (Data Preparation and EDA)**
  Clean and preprocess crime and demographic data. Aggregate by area and time, and perform exploratory analysis (maps, correlations, and trend plots).
- **October 25–November 3, 2025 (Method 1: GLM)**
  Fit Poisson/Negative Binomial GLMs to model crime counts using geographic and temporal predictors. Assess model fit and identify key linear relationships.
- **November 4–11, 2025 (Method 2: GAM)**
  Extend to semiparametric GAMs to capture nonlinear effects of predictors. Visualize smooth terms and compare performance with GLM results.
- **November 12–18, 2025 (Method 3: GAMM)**
  Introduce spatial and temporal random effects to account for neighborhood and seasonal dependence. Evaluate model improvements and check residual structure.
- **November 22, 2025 (Oral Presentation Slides Due)**
  Prepare slides summarizing data, methods, and early findings. Each team member presents one section.
- **November 23–December 5, 2025 (Final Analysis and Report)**
  Finalize all model results and comparisons (GLM vs. GAM vs. GAMM). Draft report sections in LaTeX and integrate figures, maps, and simulation findings.
- **December 6, 2025 (Soft Deadline)**
  Submit draft report for review and make necessary revisions.
- **December 7–12, 2025 (Revision and Packaging)**
  Verify reproducibility of code and figures. Prepare final archive with LaTeX, slides, and README instructions.
- **December 13, 2025, 6 PM (Final Submission)**
  Submit final report, slides, and code package on Canvas by the absolute deadline.

B References

Dataset: https://catalog.data.gov/dataset/crimes-2001-to-present