Influential Attributes on Crime Rates in Chicago (2001 - Present)

November 22, 2025

Nina Bryan (Team Lead) | Marissa Fluharty | Rachael Gu

Agenda

Background

Data Cleaning & Analysis

Method One: GLM

Method Two: GAM

Numerical Study

Conclusion

Background

Background

Research Problem: Predict crime counts to identify high-risk areas and inform resource allocation

Data Characteristics

- 8.5M reported crimes (2002-2024) with substantial spatial heterogeneity
- Count data requiring assessment for potential overdispersion
- 25 police districts covering 234 square miles
- Diverse crime types: theft, assault, narcotics, property damage, violent crimes
- Substantial variation across space, time, and context

Key Observations

- Crime is not uniformly distributed across the city
- Certain locations experience persistent high crime (hotspots)
- Temporal patterns exist: seasonal trends, monthly variation
- Crime concentrates in specific contexts (retail corridors, transit hubs, residential areas)

The Challenge: With limited police resources, how do we identify when and where to deploy officers most effectively?

Complementary Methods

GLM (Method One):

Focus: Interpretability and coefficient estimation - direct interpretation of district/crime type effects

- Parametric framework with explicit functional forms
- Categorical predictors: Crime Type, District, and Month
- Quantitative predictors: Year
- Addressed overdispersion (Pearson statistic = 14.81 originally) using quasi-Poisson
- Strengths: Direct coefficient interpretation, parsimonious structure

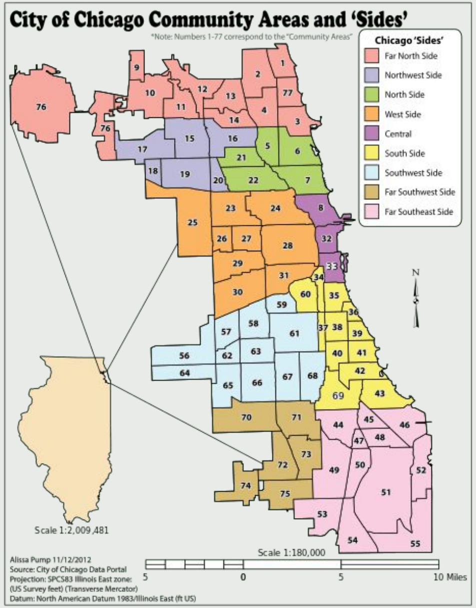
GAM (Method Two):

Focus: Pattern discovery and spatial detail

- Semi-parametric with flexible smoothing
- Continuous spatial coordinates with 2D smooth + temporal and arrest rate smooths
- Captures complex spatial patterns (e.g., O'Hare vs. Loop vs. South Side)
- Allows nonlinear effects of arrest rates, temporal trends
- Strengths: Adaptability to local geographic patterns, fine-grained spatial predictions

Data Cleaning & Analysis

Data Description



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.

Key Variables

The dataset includes the following relevant variables:

- **ID, Case Number**: Unique identifiers for each incident
- **Date:** Date and time of occurrence
- **Block:** Partially redacted address (block level only for privacy)
- **Primary Type:** Type of crime (e.g., theft, battery, burglary, assault)
- **Description:** Detailed description of the crime
- **Location Description:** Type of location where crime occurred (e.g., street, residence, parking lot, school)
- **Arrest:** Boolean indicator of whether an arrest was made
- **Domestic:** Boolean indicator of whether incident was domestic-related
- **Beat, District, Ward, Community Area:** Geographic/administrative boundaries
- **Latitude, Longitude:** Geocoded location coordinates

Data Cleaning

Handling NAs, Cleaning Each Variable, Data Manipulation

Dropped ~8% of ~8.5 million rows.

Handling NAs:

- 7 variables out of 22 with NA percentage greater than 5
- Almost all of the NA values come from the first two years of the dataset, significant amount in the first year.
- Decided to drop the first year of the dataset now focusing on 2002 - Present.
- After removing 2001, the remaining NA percentages are around 1%, so we safely drop the rest.

Cleaning Each Variable:

- Drop nonexistent districts (~0.00005%), drop district that is formed in 2012 (~0.003%)
- Drop rows with X.Coordinates and Y.Coordinates out of range (~0.001%)
- Drop nonexistent Community areas (~0.0008%)

Data Manipulation

- Gained Month, Day, Hour, Weekday from Date
- Primary.Type narrowed from 33 to 11 categories in Narrow.Type
- Location Description narrowed from 218 to 12 categories in Area.Description

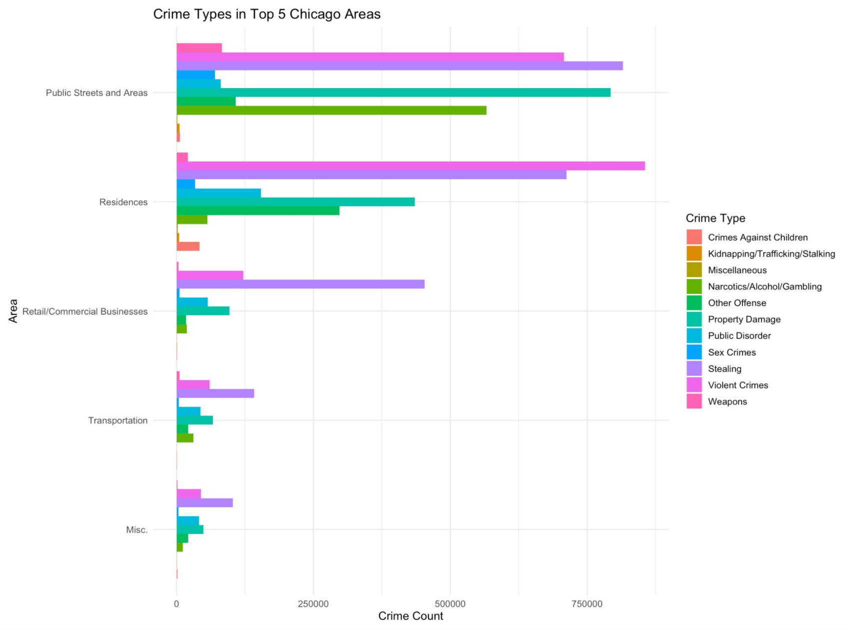
Data Analysis

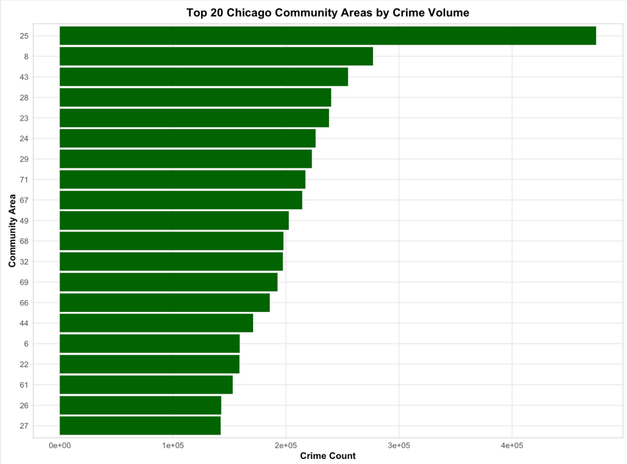
Focused on 4 Main Questions.

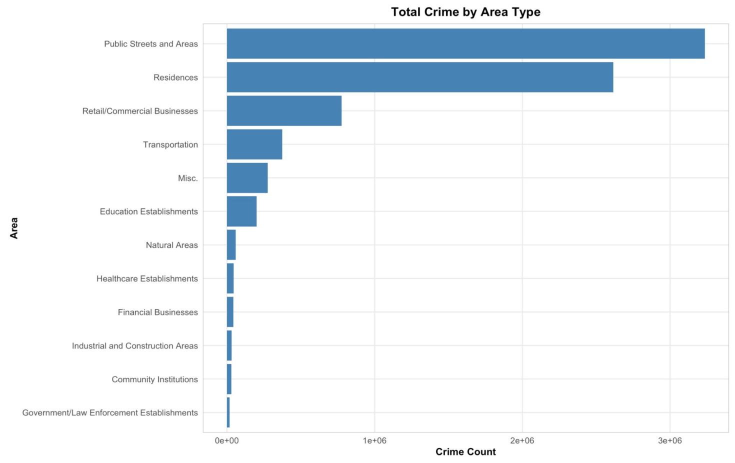
1. Which neighborhoods in Chicago experience higher rates of crime, and how does this vary across different types of crime?
2. How do crime counts change across time?
3. What is the relationship between location characteristics and crime frequency?
4. Can we identify spatio-temporal hotspots that persist or shift over time?

Question One

Which neighborhoods in Chicago experience higher rates of crime, and how does this vary across different types of crime?

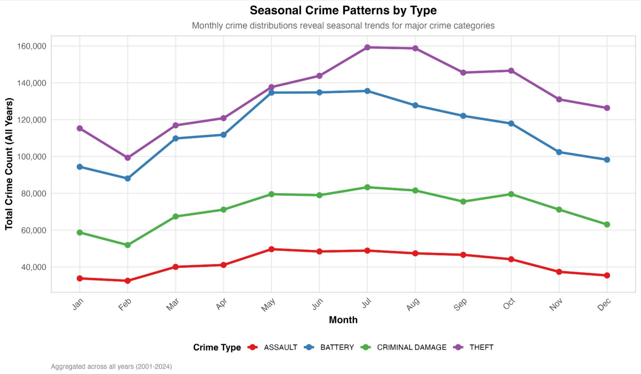


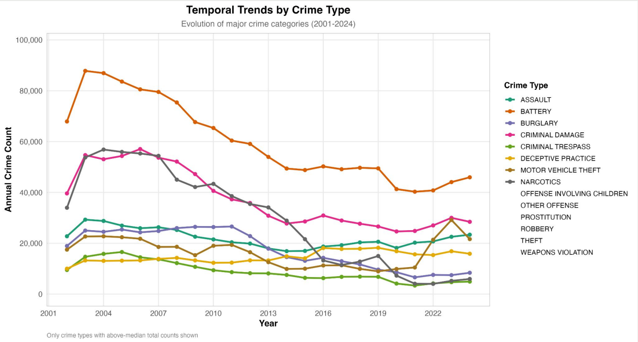


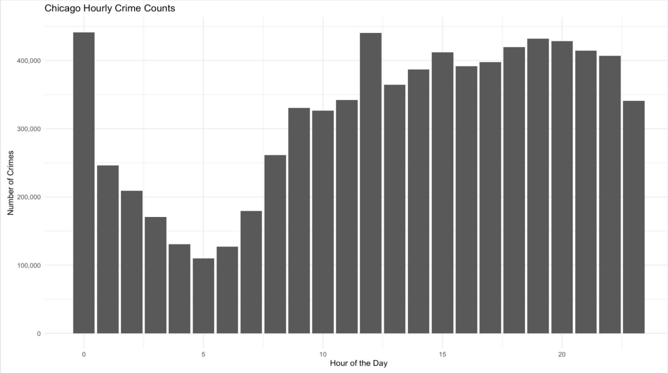


Question Two

How do crime counts change across time?

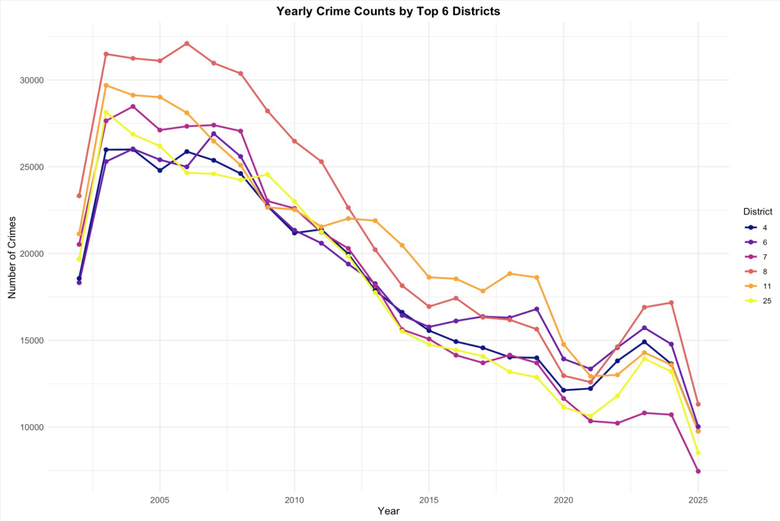


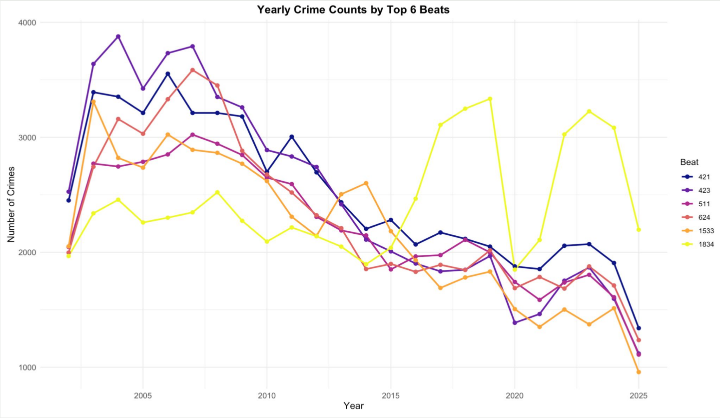


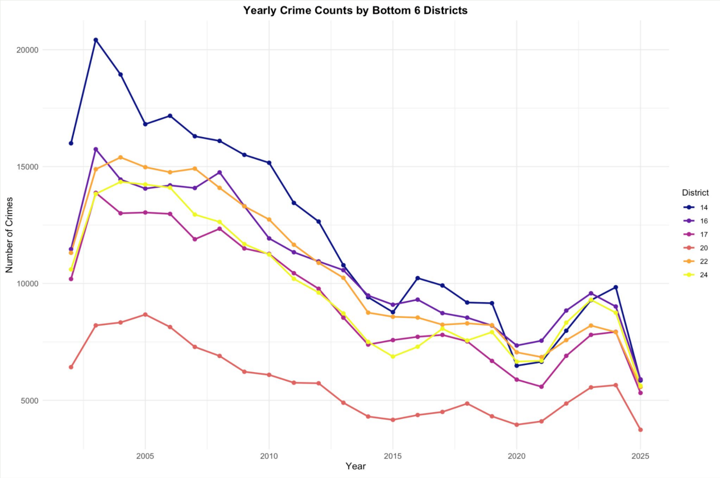


Question Three

What is the relationship between location characteristics and crime frequency?

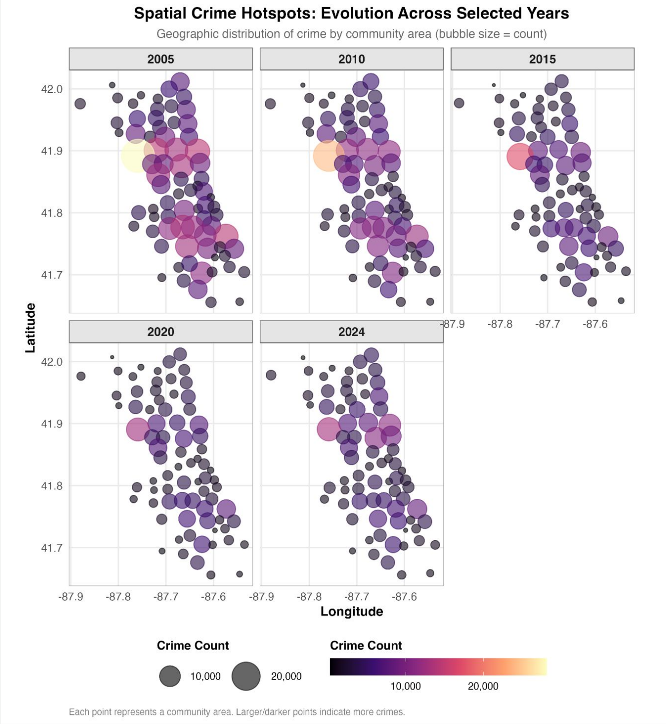


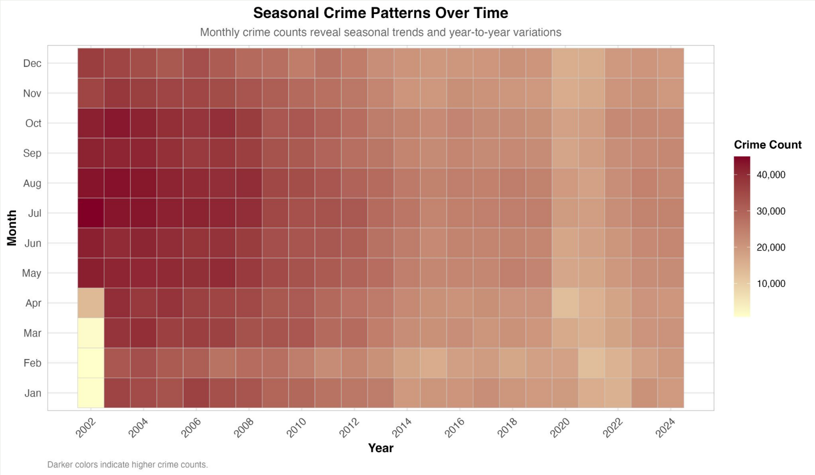


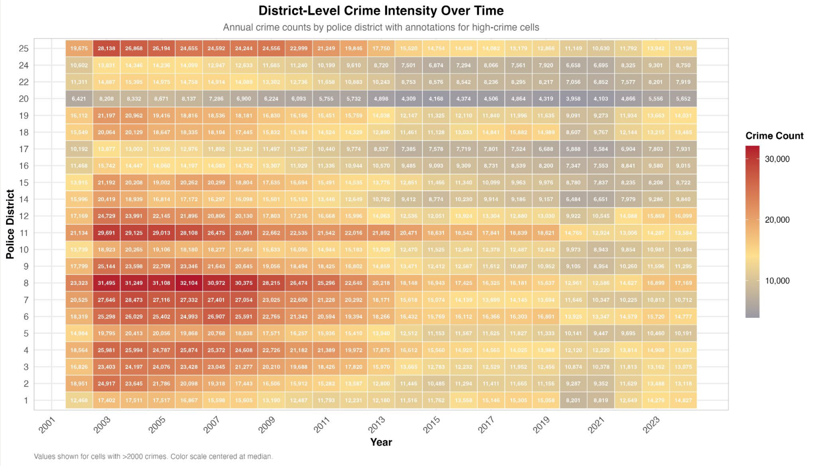


Question Four

Can we identify spatio-temporal hotspots that persist or shift over time?







Method One: GLM

Motivation for Approach

Generalized Linear Models (GLMs):

1. Suitable for count data, such as the number of crimes
2. Account for the substantial variation in crime data due to differences across locations and time

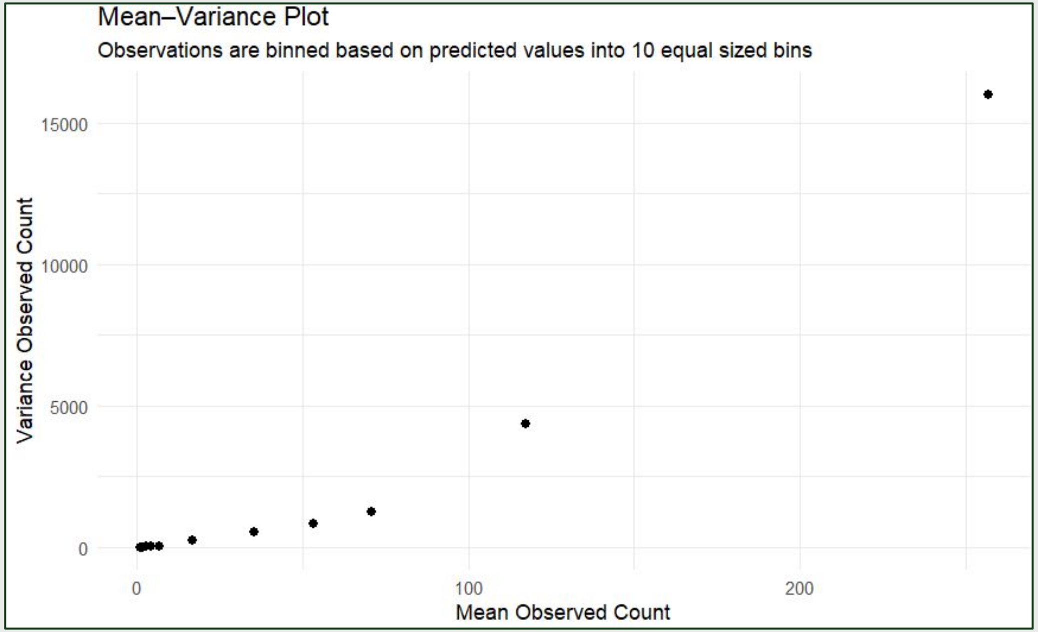
Method Steps

1. Fit initial Poisson model
    - Categorical Predictors: crime type, district, and month
    - Quantitative Predictors: year

2. Collapse groups in categorical variables
    - Month grouped by similar coefficients using k-means
    - District and crime type grouped manually for interpretability

3. Use cross-validation to choose between quasi-Poisson and negative binomial for the final model

Diagnostic Checks



Initial Poisson Model:
$\log E[\mathrm{CrimeCount}_i] = \beta_0 + \beta_1(\mathrm{CrimeType}_i) + \beta_2(\mathrm{District}_i) + \beta_3(\mathrm{Year}_i) + \beta_4(\mathrm{Month}_i)$

Mean-Variance Plot:
Variance > Mean

Pearson Dispersion Statistic:
A value of 14.81 indicates over-dispersion

Crime Type Groupings

Group Number	Group Name	Crime Types
1	Violent Crimes	ASSAULT, BATTERY, CRIMINAL SEXUAL ASSAULT, HOMICIDE, KIDNAPPING, ROBBERY, SEX OFFENSE, STALKING, INTIMIDATION
2	Property / Financial Crimes	BURGLARY, MOTOR VEHICLE THEFT, THEFT, DECEPTIVE PRACTICE, CRIMINAL DAMAGE, ARSON, CRIMINAL TRESPASS
3	Drugs / Controlled Substances	NARCOTICS, OTHER NARCOTIC VIOLATION, LIQUOR LAW VIOLATION
4	Weapons-Related	CONCEALED CARRY LICENSE VIOLATION, WEAPONS VIOLATION
5	Public Order / Miscellaneous	PUBLIC PEACE VIOLATION, PUBLIC INDECENCY, OBSCENITY, RITUALISM, GAMBLING, OTHER OFFENSE, NON-CRIMINAL, INTERFERENCE WITH PUBLIC OFFICER, PROSTITUTION
6	Human-Related Crimes	HUMAN TRAFFICKING, OFFENSE INVOLVING CHILDREN

District Groupings

Group Number	Group Name	Districts
1	Area 1	2, 3, 7, 8, 9
2	Area 2	4, 5, 6, 22
3	Area 3	1, 12, 18, 19, 20, 24
4	Area 4	10, 11, 15
5	Area 5	14, 16, 17, 25

Source:
https://chicagopd.maps.arcgis.com/apps/instant/nearby/index.html?appid=11a23d43d62b4f929dd0ec0f8c013506&sliderDistance=0.1

Year

We picked Option 2 for a simpler interpretation.

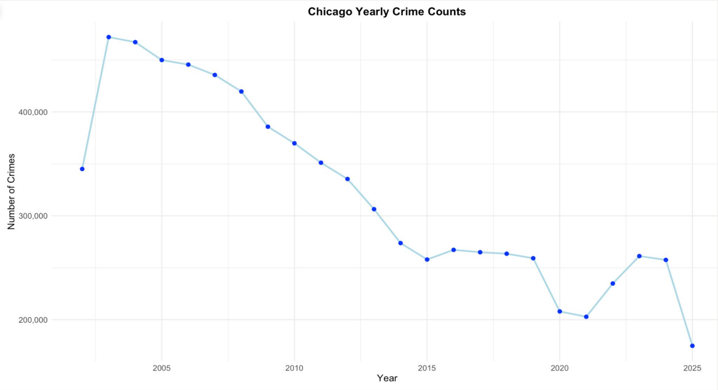
Option 1: Categorical

- Grouped by similar coefficients with k-means
- Cutoff at 2012 and 2013 is hard to interpret

Group Number	Group Name	Years
1	Early Period	2002 - 2012
2	Later Period	2013 - 2025

Option 2: Quantitative

- Slight fluctuations in crime counts year to year (especially during COVID)
- Overall, crime has been steadily decreasing
  - One coefficient for year is expected to be adequate to capture this decreasing trend



Month Groupings

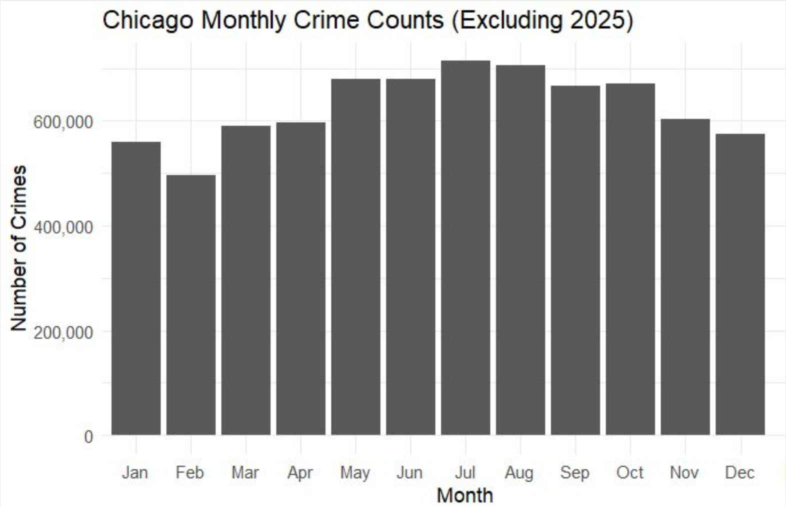
Crimes rates tend to increase during warmer months:

- More opportunities for social interaction
- School is out

Crime rates tend to decrease during colder months:

- People stay indoors

Group Number	Group Name	Years
1	Cold	November, December, January, February, March, April
2	Warm	May, June, July, August, September, October



Cross Validation

Predictors:

- **Crime Type:** violent crimes, property/financial crimes, drugs/controlled substances, weapons-related, public order/miscellaneous, human-related crimes
- **District:** area 1, 2, 3, 4, and 5
- **Year:** quantitative
- **Month:** cold and warm

Quasi-Poisson outperformed negative binomial on both RMSE and MAE, though the difference was negligible.

Fold	Model	RMSE	MAE
1	QP	83.40	53.33
2	QP	84.42	54.22
3	QP	83.37	53.50
4	QP	84.62	53.95
5	QP	82.82	53.44
1	NB	83.78	53.39
2	NB	84.82	54.32
3	NB	83.71	53.56
4	NB	85.12	54.14
5	NB	83.20	53.58

Quasi-Poisson

Reference Group:
- violent crimes, area 1, cold

Main Takeaways:
- Property and financial crimes are the most common
- Area 1 has the highest number of crimes
- Yearly crime counts show a decreasing trend
- Higher crime counts in warmer months

Term	Estimate	Exp(coef)	Interpretation
Property / Financial Crimes	0.60	1.82	Occur about 82% more often than violent crimes holding other factors constant
Drugs / Controlled Substances	0.25	1.28	About 28% higher
Weapons-Related	-1.06	0.35	About 65% lower
Public Order / Miscellaneous	-0.74	0.48	About 52% lower
Human-Related Crime	-1.78	0.17	About 83% lower
Area 2	-0.10	0.90	About 10% lower crime counts than Area 1 holding other factors constant
Area 3	-0.26	0.77	About 23% lower
Area 4	-0.08	0.92	About 8% lower
Area 5	-0.29	0.75	About 25% lower
Year	-0.03	0.97	Crime decreases by about 3% each year hold other factors constant
Warm	0.15	1.16	Crimes increases by about 16% from cold months holding other factors constant

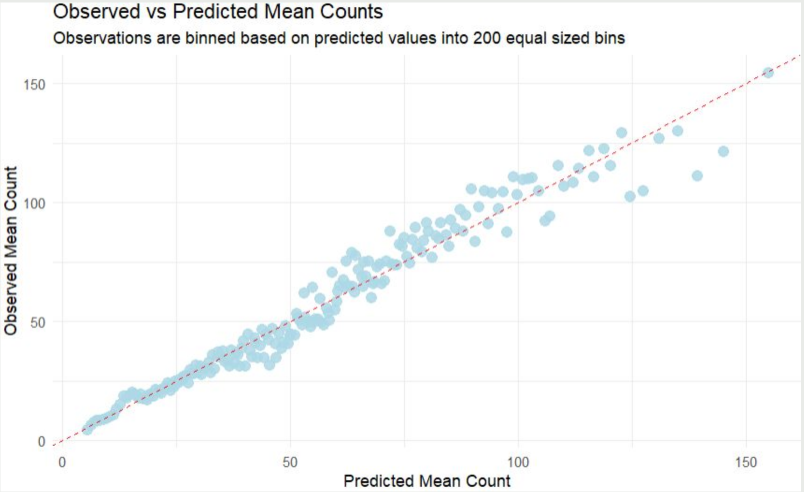
Model Evaluation

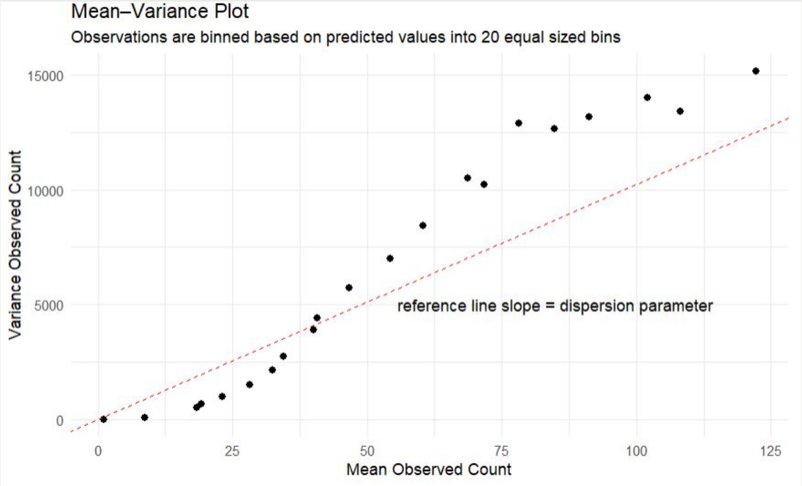
Dispersion Parameter: 102.47

MAE (fit on all data): 53.68
- Predictions differ from observed counts by ~54 crimes on average

RMSE (fit on all data): 83.72
- RMSE > MAE indicates large errors exist

Mean-Variance Relationship:
- Quasi-Poisson assumes variance grows linearly with mean
- Variance increases nonlinearly with mean





Limitations

Data:

- No socio-economic or demographic variables
- No population data per district

Model:

- Overdispersion addressed but not fully modeled

Design:

- Collapsing categorical variables loses nuance
- No interactions included
  - Certain crime types may be more common in certain months

Method Two: GAM

Motivation for Approach

1.) Crime patterns exhibit complex nonlinear relationships with space and time

2.) Traditional linear models fail to capture spatial hotspots and seasonal variations

3.) GAMs provide flexible smoothing to model these patterns without rigid assumptions

4.) Built-in spatial and temporal smoothing naturally suits geolocation crime data

5.) Handles count data with appropriate link functions (negative binomial)

Method Steps

Model <chr>	Description <chr>	N <int>	AIC <dbl>	ΔAIC <dbl>	Dev.Explained.Pct <chr>
Model 0	Null (intercept only)	50000	118060.1	4510.962	0.00%
Model 1	Spatial smooth only	50000	117678.4	4129.190	3.36%
Model 2	+ Temporal smooth	50000	117644.2	4095.017	3.66%
Model 3	+ Narrow.Type	50000	117495.8	3946.671	4.90%
Model 4	+ Area.Description	50000	117049.0	3499.806	8.38%
Model 5	+ District (RE)	50000	117042.3	3493.141	8.53%
Model 6	+ Arrest rate smooth	50000	113549.2	0.000	34.46%

7 rows

```
# Model 6: Add arrest rate smooth (FINAL MODEL)
cat("Fitting Model 6: + Arrest rate smooth (FINAL MODEL)...\n")
gam6 <- gam(Crime.Count ~
              s(Longitude, Latitude, bs = "tp", k = 50) +
              s(Time.Numeric, bs = "cr", k = 20) +
              s(Arrest, bs = "cr", k = 10) +
              s(District, bs = "re") +
              Crime_Category +
              Location_Category,
            family = nb(link = "log"),
            data = crime_gam_data,
            method = "REML")
```

- Incrementally added smooth terms to compare model fit across nested GAM structures
- Evaluated AIC and deviance explained to determine optimal model complexity
- Spatial thin plate regression spline (bs = "tp", k = 50) chosen for computational efficiency and flexibility in modeling city-wide crime patterns across 21,847 spatial locations
- Negative binomial family chosen to handle crime overdispersion
- Final Model 6 selected: lowest AIC (113,549) and highest deviance explained (34.46%)

Overdispersion Assessment

```r
# Model 6: Add arrest rate smooth (FINAL MODEL)
cat("Fitting Model 6: + Arrest rate smooth (FINAL MODEL)...\n")
gam6 <- gam(Crime.Count ~
              s(Longitude, Latitude, bs = "tp", k = 50) +
              s(Time.Numeric, bs = "cr", k = 20) +
              s(Arrest, bs = "cr", k = 10) +
              s(District, bs = "re") +
              Crime_Category +
              Location_Category,
            family = nb(link = "log"),
            data = crime_gam_data,
            method = "REML")
```

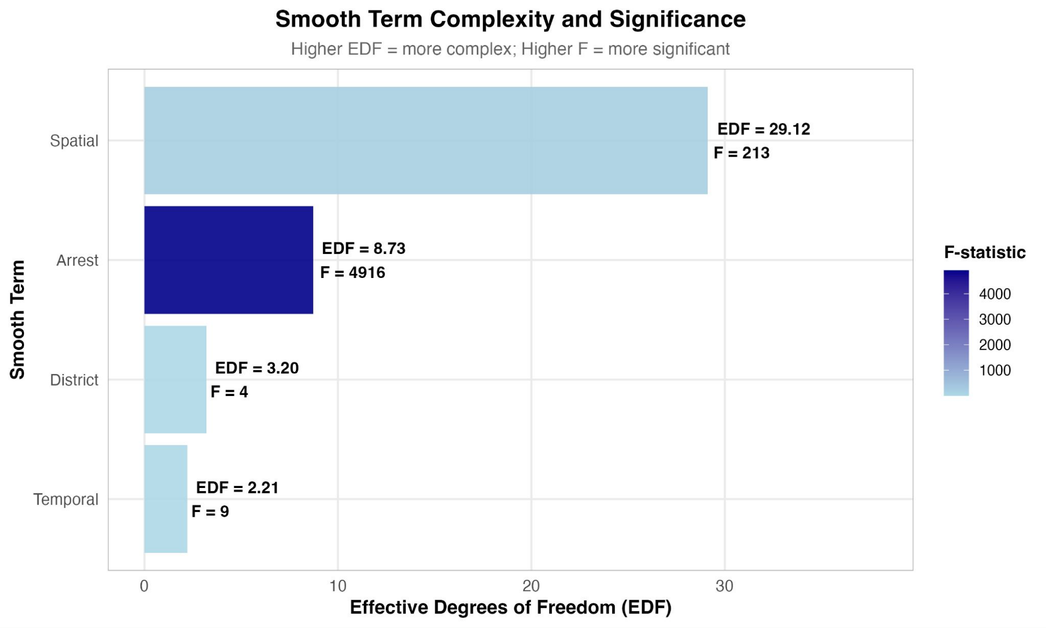
Distribution Family Selection:

- Compared Poisson and Negative Binomial
- Pearson dispersion = 0.34 after aggregation and smoothing
- Models performed equivalently (ΔAIC = 2.8)
- Retained Negative Binomial for consistency with
- criminology literature and robustness

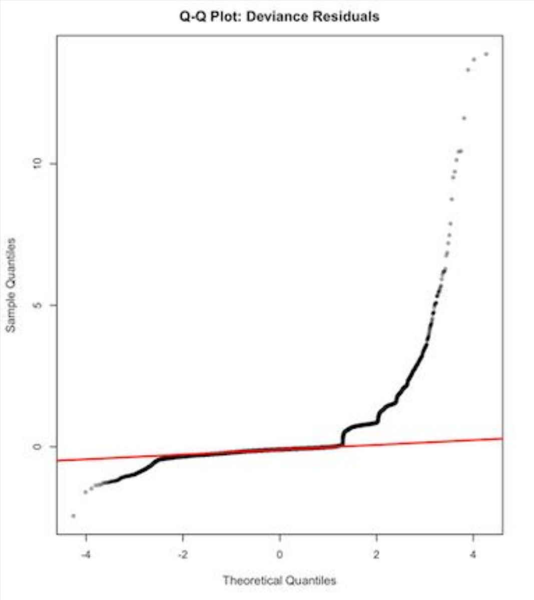
Note: Spatial aggregation and comprehensive smoothing effectively eliminated typical crime count overdispersion

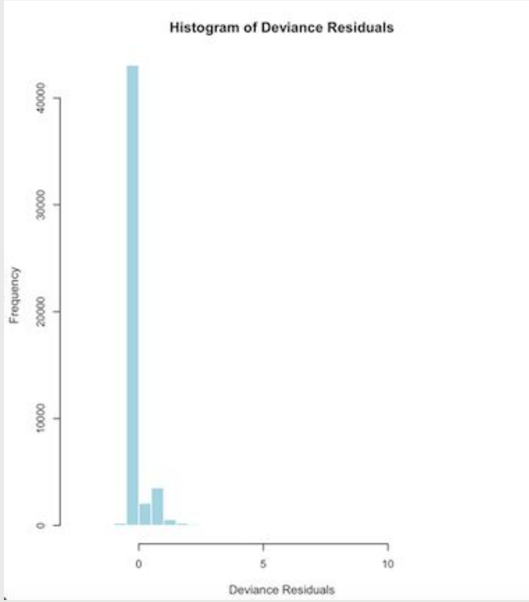
Smooth Contributions

- Spatial smooth dominates (EDF=29.12, F=213) - location is primary driver
- Arrest rate highly significant (EDF=8.73, F=4916) - strong predictor
- District and temporal effects are modest but significant
- Model explains 34.46% of deviance through smoothing alone



Diagnostic Checks





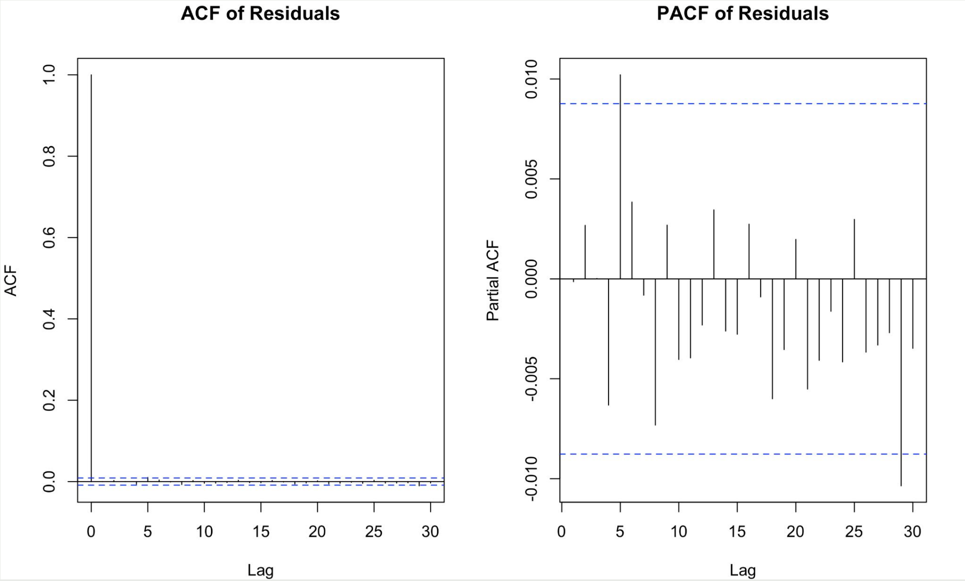
Residuals show heavy right tail typical of overdispersed crime count data

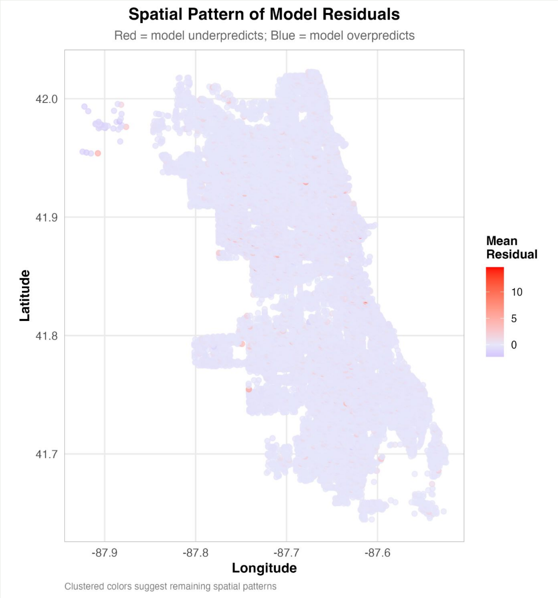
Center of Q-Q plot aligns well with theoretical quantiles, indicating good bulk fit

Deviance residuals mostly close to 0, confirming model captures average crime intensity well

Extreme residuals correspond to rare, high-count events (outliers expected in crime data)

Diagnostic Checks (Temporal + Spatial)





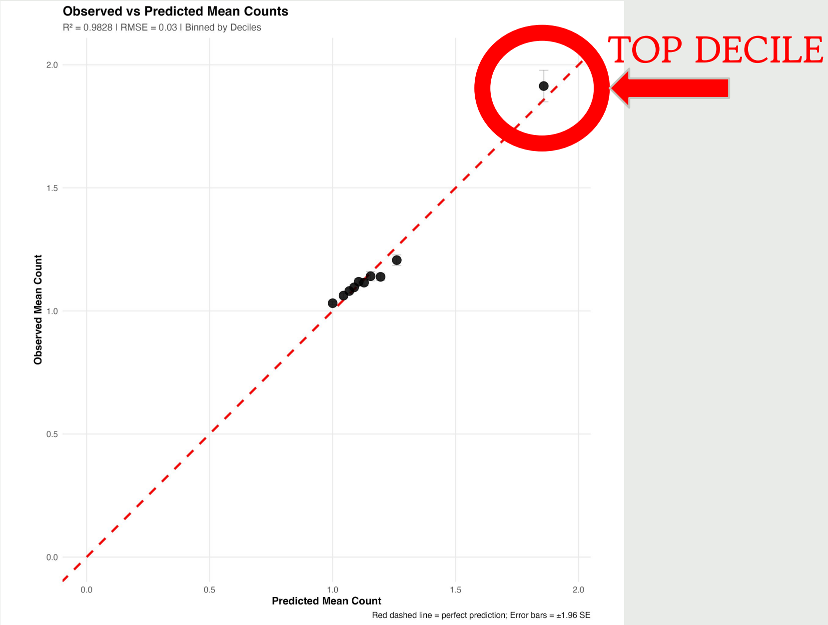
ACF shows minimal temporal autocorrelation, indicating model captured monthly seasonality

PACF spikes remain within confidence bounds, supporting lack of lingering temporal structure

Spatial plot shows no coherent residual clusters, confirming spatial smooth absorbed major patterns

Remaining residuals appear randomly scattered, consistent with a well-specified GAM

Model Findings




```
```{r}
summary(fitted(gam_final))
table(cut(fitted(gam_final), breaks = 10))
```

   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
 0.8397 1.0682 1.1163 1.1904 1.1936 17.0331

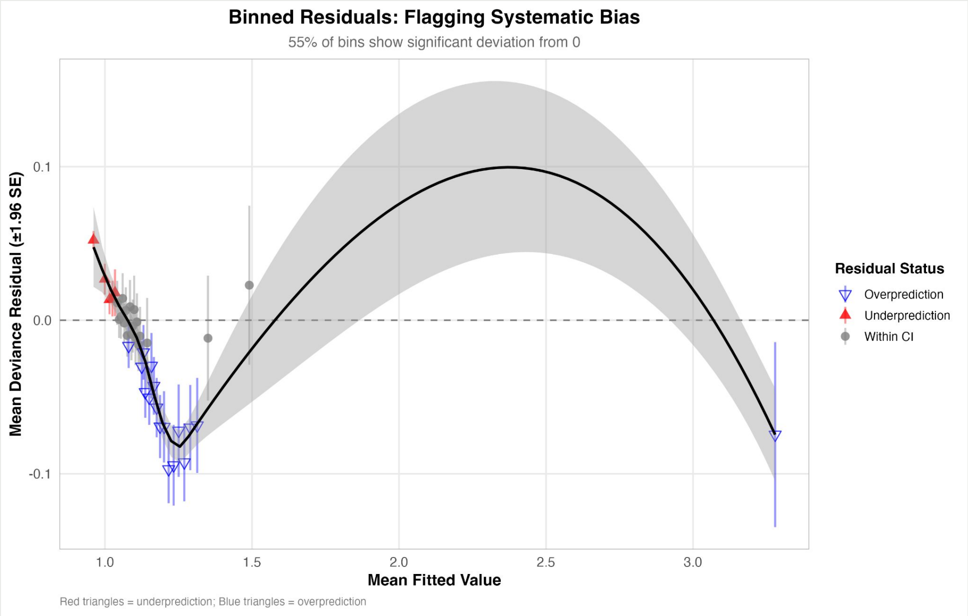
 (0.824,2.46] (2.46,4.08]   (4.08,5.7]   (5.7,7.32] (7.32,8.94] (8.94,10.6] (10.6,12.2] (12.2,13.8] (13.8,15.4]   (15.4,17]
        49365   377   158   47   22   21   9   0   0   1
```

"How well does the model predict monthly crime counts for a specific crime type in a ~2-city-block (2.3 acres) area?"

Performance Metrics:

- $R^2 = 0.9828$ - exceptional predictive accuracy
- RMSE = 0.03 - very low prediction error
- Points align tightly with 1:1 line across all count ranges
- Model well-calibrated from low-risk to high-risk areas

Diagnostic Checks



- The LOESS smoother shows a clear inverted U-shaped pattern, indicating systematic nonlinear bias across the prediction range.
- The model underpredicts at low fitted values (<1.0), achieves good calibration in a narrow range (1.1–1.2), then overpredicts in the moderate range (1.2–1.35).
- The confidence band widens substantially in the 1.5-3.0 range due to **data sparsity**—crime is spatially concentrated, with most locations having either low crime (<1.5) or hotspot activity (>3.0), and few "moderate-crime" areas.
- 55% of bins fall outside their confidence intervals, demonstrating meaningful miscalibration rather than random residual variation.
- This systematic pattern indicates the model could benefit from additional structure—such as crime-type-specific spatial smooths or nonlinear interactions—to improve calibration across the full prediction range.

Model Findings: High Avg Prediction Areas

Lat_rounded <dbl>	Long_rounded <dbl>	Latitude <dbl>	Longitude <dbl>	Mean_Predicted <dbl>	Max_Predicted <dbl>	Mean_Observed <dbl>	N_Observations <int>	Crime_Types <chr>	Location_Types <chr>
41.940	-87.651	41.94045	-87.65108	10.303936	10.303936	8.00000	1	Stealing	Retail/Commercial Businesses
41.924	-87.766	41.92398	-87.76629	10.022248	10.022248	8.00000	1	Stealing	Retail/Commercial Businesses
41.939	-87.650	41.93860	-87.64958	9.999951	9.999951	6.00000	1	Stealing	Retail/Commercial Businesses
41.879	-87.755	41.87915	-87.75502	9.797222	9.797222	9.00000	1	Stealing	Retail/Commercial Businesses
41.766	-87.664	41.76600	-87.66382	9.700002	9.739503	7.00000	2	Stealing	Retail/Commercial Businesses
41.865	-87.666	41.86469	-87.66633	9.503852	9.503852	8.00000	1	Violent Crimes	Education Establishments
41.685	-87.643	41.68507	-87.64318	9.488383	9.488383	6.00000	1	Stealing	Retail/Commercial Businesses
41.926	-87.786	41.92642	-87.78555	9.470159	9.470159	6.00000	1	Stealing	Retail/Commercial Businesses
41.743	-87.634	41.74271	-87.63409	9.162242	10.371414	19.33333	3	Stealing	Retail/Commercial Businesses
41.910	-87.743	41.90966	-87.74271	8.955630	8.955630	9.00000	1	Stealing	Retail/Commercial Businesses

1-10 of 10 rows

The top hotspot is at:

Latitude: 41.940, Longitude: -87.651
Mean Predicted: 10.30
Mean Observed: 8.0
Location: Retail/Commercial Businesses
Crime Type: Stealing

This location is near West Belmont Avenue, Lakeview

Observation with High Avg Observed Crime Count

(Latitude: 41.743 , Longitude: -87.634) = West 83rd Street area (Southside of Chicago)

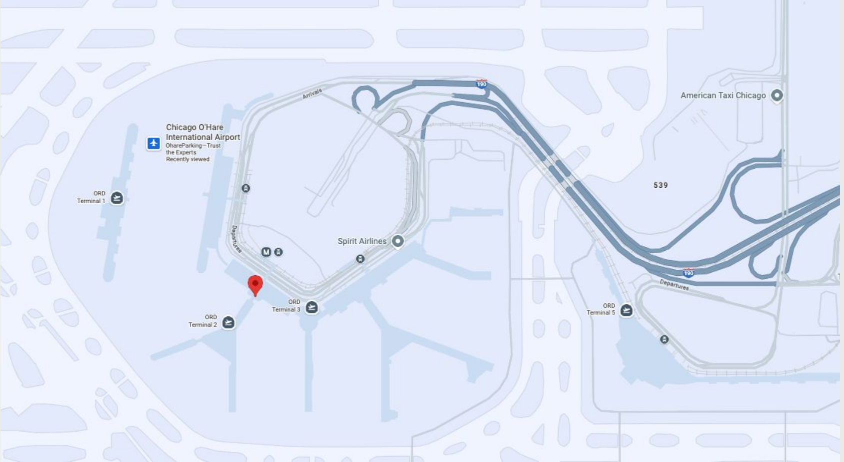
Model Failure Location (Underprediction)

The model significantly underpredicted here

- Predicted ~9 incidents
- Actually had ~19 incidents
- Model missed by more than 100%
- Observed Count more than **double** the predicted Count
- Outlier location suggests **locally sharp crime spikes** not captured by smooth spatial terms

Observation with Highest Predicted Crime Count

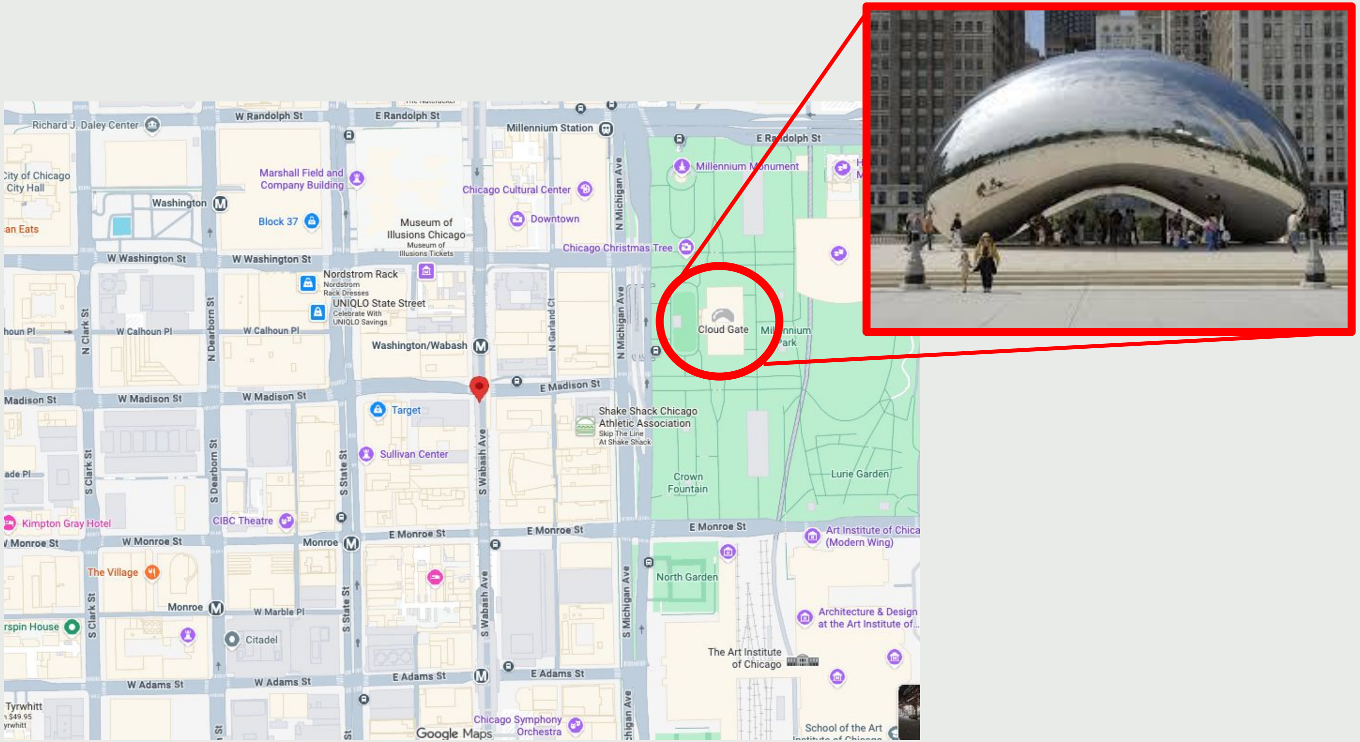
(Latitude: 41.976 Longitude: -87.905) = O'Hare International Airport (District 16)



- Significance: Model correctly identifies airport-specific crime patterns distinct from street-level crime
- Transportation offenses include TSA violations, security breaches, and federal jurisdiction incidents
- Prediction 70% higher than typical retail hotspots—demonstrates model learns location-type-specific patterns
- Post-9/11 context (2003): heightened security enforcement likely influenced incident rates

Latitude <dbl>	Longitude <dbl>	Lat_rounded <dbl>	Long_rounded <dbl>	Year <dbl>	MonthNum <dbl>	Crime_Category <fctr>	Location_Category <fctr>	District <fctr>	Predicted <dbl>	Observed <int>
41.97629	-87.90523	41.976	-87.905	2003	3	Other Offense	Transportation	16	17.03314	8

Honorable Mentions (High Prediction Count)



(Latitude: 41.881, Longitude: -87.626) = The Loop; Crime: Stealing

GAM Limitations

- **Computational intensity** – Fitting flexible smooths with 50,000 observations and k=50 basis functions requires significant computational resources compared to parametric GLM
- **Interpretability trade-off** – Smooth terms require visualization and are less intuitive than direct GLM coefficients; difficult to summarize spatial effects in a single number
- **Extrapolation risk** – Model should not be used to predict crime outside observed spatial domain (beyond Chicago boundaries) or temporal range (pre-2002 or far future)
- **Spatial aggregation dependency** – Results depend on choice of 111m × 83m grid resolution; finer (address-level) or coarser (district-level) aggregation may reveal different patterns

GAM Limitations

- **Associational, not causal** – Model identifies correlations and patterns but cannot establish causation; observed relationships (e.g., arrest rates and crime) do not imply intervention effects without experimental or quasi-experimental design

- **Smooth relationship assumptions** – Penalized smooths favor gradual changes and may underestimate sharp spatial boundaries (e.g., crime "cliffs" at district borders or infrastructure barriers)

- **Geographic heterogeneity** – Model performs better in well-represented areas (downtown, North Side) than undersampled neighborhoods (some South Side locations show systematic underprediction)

Numerical Study

Numerical Simulation

Create 3 scenarios to study:

- Scenario One: Crime changes linearly over time.
- Scenario Two: Crime changes non-linearly over time.
- Scenario Three: Crime changes mildly non-linearly over time.

For time increments $t = 1$ to $T = 300$ we will create the behavior from each scenario of the mean crime count over time:

1. $m_1(t) = \exp\{2 - 0.01t\}$
2. $m_2(t) = \exp\{2 - 0.01t + 0.4\sin(0.04t)\}$
3. $m_3(t) = \exp\{2 - 0.01t + 0.2\sin(0.04t)\}$

Now for each t value we have a corresponding mean value for each scenario.

To get the y value we can draw a random variable from the Poisson distribution.

1. $Y_t \sim \text{Poisson}(m_1(t))$
2. $Y_t \sim \text{Poisson}(m_2(t))$
3. $Y_t \sim \text{Poisson}(m_3(t))$

We then fit the GLM and GAM models on this y value with the following formulas:

- glm_model = glm(y ~ t, family = poisson)
- gam_model = gam(y ~ s(t), family = poisson)

Calculate the corresponding MSE's for each model and scenario (6 MSE values)

We will repeat the above for N = 500 repetitions then average all of the MSE values.

Key Findings, Highlights and Improvements

Key Findings

- When the true relationship is linear GLM performs best.
- When the true relationship is nonlinear GAM performs best.
- When the true relationships is mildly nonlinear GAM performs best.

Highlights

- GAM performs best consistently when there is a nonlinear relationship.
    - If the true distribution of the data is unknown GAM is the safest option, unlikely that strict linearity is adhered to.

Improvements - How does it apply to our data?

- Utilize GLM only if the EDA shows true linearity with crime count over time.
- Use GAM unless strong evidence is present to shift towards GLM.
- The simulation confirms support for a GAM model.
- It is unlikely that the GLM outperforms the GAM model unless there is true linearity.

Scenario	MSE for GLM	MSE for GAM
1	0.0198	0.0277
2	1.1376	0.0926
3	0.1212	0.0726

Conclusion

Method Comparisons

Feature	**GLM (Method 1)**	**GAM (Method 2)**
Spatial Representation	Categorical (District 1-25)	Continuous (Lat/Long grid)
Spatial Resolution	25 districts (grouped into 5 areas)	21, 847 cells (111m x 83m)
Temporal Representation	Categorical (grouped month) and Quantitative (year)	Continuous smooth
Functional Form	Linear (log scale)	Flexible smooths (learned)
Overdispersion	Quasi-Poisson or NB	Negative Binomial
Model Complexity	Lower (fewer parameters)	Higher (smooth terms)
Interpretability	Direct coefficients	Visual smooth effects
Predictions	District-level Averages	Neighborhood-specific
Best For	City-wide policy	Targeted interventions

Thank You!!

Nina Bryan (Team Lead) | Marissa Fluharty | Rachael Gu

References

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

Numerical Simulation: Full title: Flexible modeling using Generalized Additive Models ...eScholarship@McGillhttps://escholarship.mcgill.ca › downloads