UNITED STATES MILITARY ACADEMY

Final Project

MA478 GENERALIZED LINEAR MODELS

SECTION H2

COL NICHOLAS CLARK

BY

CADET TOBIAS HILD '24, CO G1

WEST POINT, NEW YORK

7 MAY 2024

X OUR DOCUMENTATION IDENTIFIES ALL SOURCES USED AND ASSISTANCE RECEIVED IN COMPLETING THIS ASSIGNMENT.

___ WE DID NOT USE ANY SOURCES OR ASSISTANCE REQUIRING DOCUMENTATION IN COMPLETING THIS ASSIGNMENT.

SIGNATURE: /s/ Tobias Hild

1 Abstract

This paper investigates the factors influencing burglary rates in Chicago, Illinois, using a dataset spanning from 2010 to 2015. By examining various demographic and economic variables alongside temporal and spatial patterns, the study aims to provide insights into the association between social and economic factors. Drawing upon existing literature on spatial and temporal patterns of criminal behavior, the paper employs generalized linear mixed-effects models to analyze the impact of different variables on burglary rates. Temporal analysis techniques including trend analysis and periodicity analysis are utilized to understand the overall direction and recurring patterns of burglaries over time. The results indicate several significant findings. Firstly, there is a general downward trend in burglary rates over the study period, with increased activity during the hotter summer months. Additionally, there is a strong neighborhood effect, suggesting that neighboring census blocks tend to exhibit similar burglary rates. Demographic variables such as wealth per capita and the proportion of young males are found to be associated with burglary rates, while unemployment shows no significant effect. Model comparison using deviance information criterion (DIC) suggests that a model incorporating neighborhood effects and an AR1 autocorrelation for a random effect by months performs best for predicting burglary rates.

2 Introduction

The understanding of the causes of crime is vital to the functioning of a city. By understanding what variables are associated with an increase in burglaries, the leaders of communities are able to address possible underlying causes of criminal activity as well as proactively allocate resources to prevent and respond quickly to crime. The goal of our project is understand the impact of a variety of potential variables on burglaries in Chicago, Illinois by utilizing modeling the number of burglaries over time and space. We utilize population data, wealth, unemployment rate, proportion of the population that is young males, temperature to represent a cyclical pattern repeating every year, and date as explanatory variable. We hope to understand which demographic and economic variables explain burglary rates in Chicago, and trends over time and space.

2.1 Literature Review

From a meta-analysis of different spatial and temporal patterns of criminal behavior conducted by Leong and Sung (2015), spatio-temporal patterns of criminal activity can be broken into several categories. Hotspots (Eck et al, 2005) are a single point of higher than usual activity. These can be modeled with a categorical variable indicating whether an area is part of the hotspot or not. Nearest neighbor models and other techniques that produce clusters are also useful for understanding which areas are part of a hotspot. In our analysis we will use a random error structure as well as neighborhood effect models to understand the geographic dispersion of our data. Temporal analysis techniques are generally more complicated since

they deal with time series data. These techniques can be categorized (Han and Kamber, 2000) as trend analysis, similarity search, periodicity analysis and sequential mining. Trend analysis deals with whether the incidence rate of crime generally increases or decreases over a period of time, long term cyclic movements, seasonal movements, and random movements. We will use the date as an explanatory variable to understand the overall direction of the change in burglaries over time as well a temperature to understand the seasonal patterns. Similarity search and periodicity analysis deal with patterns that are repeated in time series data, where periodicity analysis deals with the time scale of the recurrance and similarity search focuses on identifying repeated patterns. More complicated methods of combining temporal and spatial analysis are also available, such as finding hotspots which appear during certain parts of a cyclic pattern, or assigning topological patterns to different parts of a time series. This paper will not focus on assigning detailed patterns to our data, but instead try to understand the overall trend of burglaries and seasonal patterns that might generalize to different times.

Temperature has been shown to be a useful covariate for predicting crime rates not only on a year to year time scale but also across time of day (Cohen and Gonzalez, 2024). Some of this variation can be explained as an association with alcohol use as a mediator between temperature and burglaries, but even when controlling for these factors a significant explained correlation between temperature and crime remains.

3 Data

In this project we are explaining Number of burglaries in month i in census block j of Chicago (denoted by y_{ij}). Our data contains observations from the years 2010 through 2015 (a total of 72 months). For each census block there is demographic information: the population of the census block, the number of young males in the census block, a standardized metric of the total wealth of that area and the unemployment proportion. All of this data is derived from the 2010 census and is constant over time. The census block represents the smallest geographic unit for which this census data is reported. In our dataset, the population of blocks varies from 12-3000, but generally about 1000 people. We standardize all of the explanatory variables to be in terms of the population of the given block; we are left with young males proportion, unemployment proportion and wealth per capita. We will model the number of burglaries using an offset for population, so our outcome of interest is burglaries per capita in a given month.

A number of variables not present in our dataset might vary across space and time, and cause a difference in the number of burglaries at a given point. In particular changes in the demographic variables in our dataset are not accounted for. Neither are changes in law and law enforcement policy over time. These and other variables will not in our dataset will be captured using random effects for both time and space. Table 1 is an overview of the sources of variation in our dataset.

Plotting the aggregated total number of burglaries over time we can see that the number of burglaries over times appears to follow a cyclical pattern with a period approximately

Sources of Variation Diagram		
Observational Units	**Sources of Explained Variation**	**Sources of Unexplained Variation**
Census Blocks within Chicago for each month from 2010-2015	Census Block location Time of year Young Male Proportion Time since Jan 2010 Wealth Unemployment	Other demographic effects Geographic variation in police Variation in Laws Willingness to Report Housing types Security technology and access

Table 1: Sources of variation in Chicago Census blocks

corresponding to a year, and decreasing over time. This suggests that the number of burglaries over consecutive months are correlated which could be captured using an AR1 error structure for a time based random variables, but also that there is some variation over time (cyclical and overall linear) that could be captured using a covariate for time. A possible covariate for the cyclical pattern we see over time is temperature. As discussed by Cohen and Gonzalez, there is some evidence that temperature is associated with incidence of crime, as well as plausible causal pathways for this relationship. In Figure 1 we see that average monthly temperature does appear related to the total number of burglaries across Chicago.

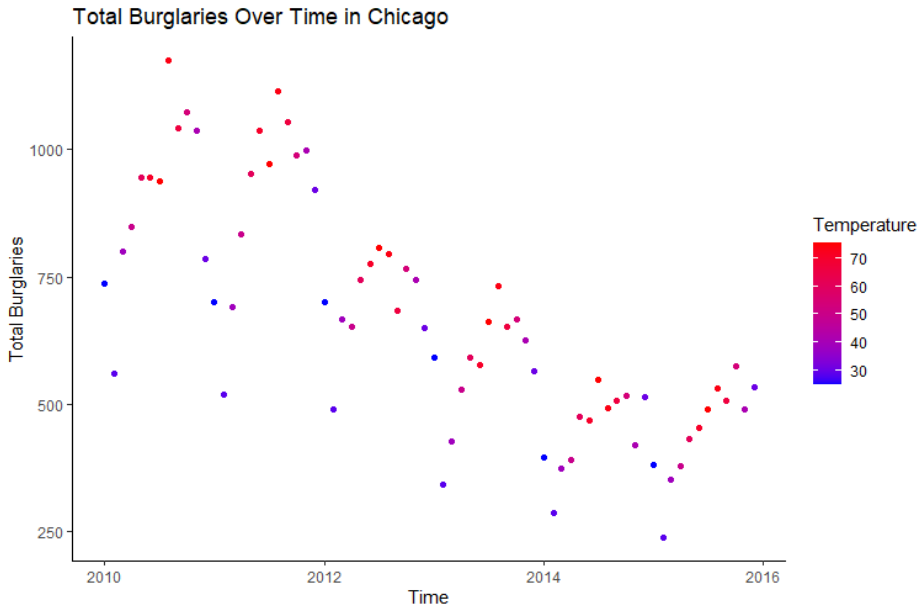

Figure 1: Aggregated Burglaries over Time

Examining the aggregated number of burglaries over census block we see that there is significant variation between different census blocks as well as geographically in areas that

are not captured in these blocks. For example at the north of our study area we see an area of relatively low aggregated burglaries, whereas to the east we see relatively more burglaries overall. This suggests that there is a greater correlation between neighboring blocks than between non-adjacent blocks. We could account for this variation by correlating adjacent blocks, or by using larger geographic areas such as counties.

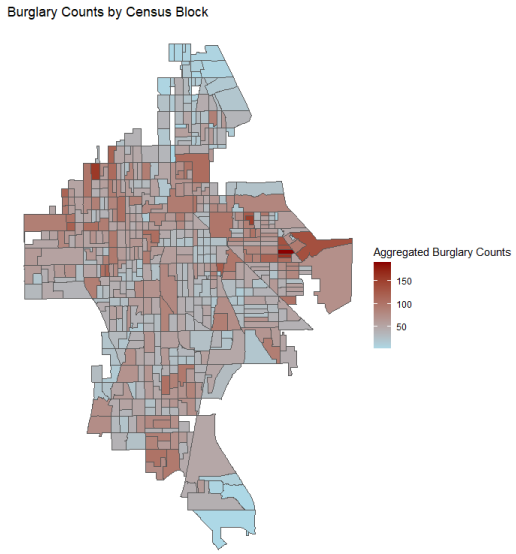

Figure 2: Aggregated Burglaries over Space

4 Methodology

We will use generalize linear mixed effects models to answer the following questions which appear to be answerable using the data at hand.

- Can we substantiate the temporal patterns from our data exploration? We will use models with time represented by monthly average temperature to determine if there is a cyclic yearly trend. We will also use months since the beginning of the time for which data is available to determine if the overall downward trend in the number of burglaries over time apparent in Figure 1 is significant. We will also fit a model using an AR1 correlation pattern for time to determine whether the patterns over time can be efficiently capture using only the information from the previous month.

- Are there spatial patterns in the number of burglaries across Chicago? We will test this hypothesis by comparing models using a BYM error structure for census blocks and a model using an uncorrelated error structure for census blocks. By comparing the precision for both of these error terms, we will be able to determine whether the variation in burglaries over census blocks follow spatial patterns or are completely at random.

- We are also interested in determining which social and economic data explain burglaries well. This may be of interesting in developing predictive models in the future. We will compare models with different combinations of explanatory variables and select those variables which are consistently significant and improve the models.

- Finally we will determine our best model. We will compare models using deviance information criterion (DIC) as well as interpretability, and residual patterns.

All of our models will be fitted using integrated nested Laplace approximation using the INLA package in R.

4.1 Models

In our first model we will try to establish trends over time. We include all demographic variables, adjusted for population. Trends over time are captured by covariates for the number of months since January 2010 (the beginning of the period for which we have data) and the mean temperature of the month. The month number is added to understand the linear trend over time. We expect the coefficient for this variable to represent the direction of the trend in burglaries over time. The coefficient for temperature should represent the direction of the cyclical trend, that is whether there are generally more burglaries in the hotter or the colder months. In this model we use a random effect for census block; this random effect captures the unique aspects of each census block assuming that each census block is independent of every other census block. Using a random effect for location allows us to control for location effects not included in our fixed effects and generalize to locations not in our dataset.

Model 1. Time Model

$$y_{ij} \sim \text{Po}(\lambda_{ij})$$

$$\log(\lambda_{ij}) = \eta_{ij}$$

$$\eta_{ij} = \beta_0 + \log(\text{Pop}_j) + \beta X + u_i$$

$$X = \begin{bmatrix} 1 \\ \text{Unemployment Proportion}_j \\ \text{Wealth Per Capita}_j \\ \text{Young Male Proportion}_j \\ \text{Months since Jan 2010}_i \\ \text{Temperature}_i \end{bmatrix} \hat{\beta} = \begin{bmatrix} (-7.021, -6.803) \\ (-0.201, 1.055) \\ (-23.455, -14.994) \\ (0.074, 2.498) \\ (-0.013, -0.013) \\ (0.007, 0.008) \end{bmatrix}$$

$u_i \sim N(0, \sigma_u^2)$ is a random component for Census Block

$$\text{Prior}(\frac{1}{\sigma^2}) \sim \text{Loggamma}(1, 10^{-5})$$

$$\frac{1}{\hat{\sigma}^2} = (3.08, 3.99)$$

We can see that our hypotheses about the time trends of burglaries are substantiated by this model. The significant negative coefficient for months since beginning of the observation period indicates a general downward trend in the number of burglaries. The significant positive coefficient for temperature indicates more burglaries occur during the summer months. We also see that a greater young male proportion tends to increase the burglary rate and wealth per capita is associated with a lower burglary rate.

The relative size of the coefficients are not necessarily representative of greater importance since the variables are not standardized.

Our second models is designed to determine whether there is a neighborhood effect; that is whether the burglary rate of a certain block helps to explain the burglary rates of adjacent blocks. We use a BYM model to examine this hypothesis. The BYM model correlates every census block to those that are adjacent to it (share a border with a given block). By fitting the model we obtain an estimate of the precision of that correlation. Higher precision indicates a stronger neighborhood effect. We use a similar technique for variation across time. The AR1 autocorrelation structure uses the previous value in a time series to predict the next value in the series. Fitting this model gives a correlation coefficient, with higher values indicating the consecutive

Model 2. Neighborhood Model

$$y_{ij} \sim \text{Po}(\lambda_{ij})$$

$$\log(\lambda_{ij}) = \eta_{ij}$$

$$\eta_{ij} = \beta_0 + \log(\text{Pop}_j) + \beta X + \gamma_j + u_t$$

$$X = \begin{bmatrix} 1 \\ \text{Unemployment Proportion}_j \\ \text{Wealth Per Capita}_j \\ \text{Young Male Proportion}_j \end{bmatrix} \hat{\beta} = \begin{bmatrix} (-7.315, -6.675) \\ (-0.218, 1.048) \\ (-23.642, -15.186) \\ (0.004, 2.435) \end{bmatrix}$$

$$\text{Precision for Block (iid component)} = (3.024, 4.009)$$

$$\text{Precision for Block (spatial component)} = (18.389, 199.561)$$

$$\text{Precision for month number} = (3.806, 17.289)$$

$$u_t = \Phi u_{t-1} + \upsilon_t$$

$$\Phi = 0.866$$

We see very similar coefficients for the demographic covariates as in model 1. This is logical since these covariates are constant over time and therefore are not effected by the change in representation of time between the models. The high value of Φ indicates that burglary rates in consecutive months tend to be very similar; although this model does not

tell us anything about the trend, in combination with Model 1 these results allow us to assert that the trends we observed earlier are generally consistent over time.

The high precision for the spatial component relative to the i.i.d component indicates a strong neighborhood effect, substantiating the hypothesis that there is some association between burglaries in adjacent blocks. We can also conclude that there is some effect for location on a larger than block scale which has an association with burglary rates. This is not accounted for in the fixed effects available in the models.

The third model we fit combines attributes of the first two models. We include fixed effects for both linear time represented by the month number and cyclical time patterns. We also removed the covariate for unemployment proportion, since this variable was not significant in the previous models (we also fit this model with unemployment proportion and found no difference in DIC). We use the BYM model for spatial variability.

Model 3. Combined Model

$$y_{ij} \sim \text{Po}(\lambda_{ij})$$

$$\log(\lambda_{ij}) = \eta_{ij}$$

$$\eta_{ij} = \beta_0 + \log(\text{Population}_j) + \beta X + u_t$$

$$X = \begin{bmatrix} 1 \\ \text{Wealth Per Capita}_j \\ \text{Young Male Proportion}_j \\ \text{Months since Jan 2010}_i \\ \text{Temperature}_i \end{bmatrix} \hat{\beta} = \begin{bmatrix} (-6.994, -6.781) \\ (-23.294, -14.859) \\ (0.145, 2.558) \\ (-0.013, -0.012) \\ (0.007, 0.008) \end{bmatrix}$$

$$\text{Precision for Block (iid component)} = (3.05, 3.96)$$

$$\text{Precision for Block (spatial component)} = (118.86, 8023.20)$$

The inferences drawn from Model 3 are the same as for the previous models. This model does not add anything to our understanding of the association between our explanatory variables and the observed burglary rate.

5 Discussion

5.1 Model Selection

We examined all three of our models and found there to be no noticeable patterns in the standardized residuals. We did not find any model which passed a goodness of fit test. Of the three models we tested, Model 2 had the lowest DIC. We recommend this model for prediction since we believe that it is more generalizable as to make a prediction for a new observation it is only necessary the data for the surrounding blocks and the preceding months; no observations of trends over time are necessary. Therefore this model has greater potential to be applied to other cities and other times.

5.2 Ethical Considerations

The results of this project are very limited and should not be used to infer the causes of any criminal activity individually or the reason for a trend. Our results address the association between various demographic variables and burglary rates but not the causes of these demographic factors or what the causal pathways between social and economic composition of a location and the burglaries committed there. We also believe that any policy made to address criminal activity should consider more than the data and the conclusions that are presented in this paper.

5.3 Conclusions

Our models give us answers to the research questions we proposed in our methodology sections. We found that there was a general downward trend in burglary rates over time in our dataset. We also found increases in burglary rates during the hotter summer months, consistent across the years included in the dataset. We also found that burglary rates tended to be highly correlated with the burglary rates from the previous month. We also found that there is a strong neighborhood effect for burglary rates; neighboring census blocks tend to have similar burglary rates. Of the demographic variables we analyzed, we found that a higher wealth per capita is associated with lower burglary rates and that a higher proportion of young males in a census block is associated with higher burglary rates. Unemployment was consistently not significant in our analysis.

6 References

ChatGPT was used to write a first draft of the abstract for this paper. We also used ChatGPT to write the code for creating the figures in this paper as well as for data wrangling. The conversation is available.

Cohen, François, and Fidel Gonzalez. 2024. "Understanding the Link between Temperature and Crime." American Economic Journal: Economic Policy, 16 (2): 480-514.

Eck, J., Chainey, S., Cameron, J. and Wilson, R. 2005, "Mapping crime: Understanding hotspots", National Institute of Justice, Washington DC.

Han, J., Pei, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U. and Hsu, M. 2001, "Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth", proceedings of the 17th international conference on data engineering, pp. 215.

Leong, Kelvin and Sung, Anna. 2015. "A review of spatio-temporal pattern analysis approaches on crime analysis." International E-Jounral of Criminal Sciences, Article 1 No 9.

6.1 Additional Materials

Code used in this project, as well as the original data are available on GitHub