UNITED STATES MILITARY ACADEMY

Spatial and Temporal Analysis of Chicago Burglaries

MA478 GENERALIZED LINEAR MODELS

SECTION H2

COL NICHOLAS CLARK

BY

CADET JACOB HYATT '25, CO C2

WEST POINT, NEW YORK

7 MAY 2024

JH MY DOCUMENTATION IDENTIFIES ALL SOURCES USED AND ASSISTANCE RECEIVED IN COMPLETING THIS ASSIGNMENT.

____ I DID NOT USE ANY SOURCES OR ASSISTANCE REQUIRING DOCUMENTATION IN COMPLETING THIS ASSIGNMENT.

SIGNATURE: HYATT.JACOB.ELIJAH.1604361520 Digitally signed by HYATT.JACOB.ELIJAH.1604361520 Date: 2024.05.07 23:48:16 -04'00'

Spatial and Temporal Analysis of Chicago Burglaries

CDT Jacob Hyatt

May 8, 2024

1 Abstract

The study investigates burglary patterns in Chicago from 2010 to 2015, aiming to determine temporal and spatial trends. The research explores correlations between burglaries per census block and covariates like population, wealth, and unemployment rate. The three models were created using INLA. Through data analysis we see a general trend of burglaries decreasing over the time period, however, there are peaks during the summer time. We also see that there are clusters of high burglaries and clusters of low burglaries. After fitting all of the models, we see that the young male proportion has a positive correlation with burglaries while the wealth proportion has a negative correlation. Model 2, incorporating spatial effects with a BYM model and AR1 for time, demonstrates the strongest performance out of the three fitted models. It highlights clustered spatial patterns and temporal correlations between consecutive months. This study sheds some light on some factors that explain burglary. Limitations include the dataset's applicability to the modern day and the complexity of crime behavior. Future research could address evolving crime dynamics and enhance model robustness by including other types of crime. Ethical considerations include the application of the model and data collection.

2 Key Words

Spatial Analysis, Temporal Analysis, Burglary, INLA, Poisson

3 Introduction

Understanding patterns and variations in burglaries in a large city helps inform policy makers and law enforcement. The goal of our project is to capture the impact of a variety of potential variables on burglaries in Chicago, Illinois by utilizing various covariates over time and space. We utilize population data, wealth, unemployment rate, young males percentage, temperature, and time. All of this data spans the months from 2010 to 2015. We have a couple of research questions to focus on. Can we substantiate the temporal patterns from our data exploration by month or by year? What are some locational patterns? Are there neighborhood patterns? Which economic and demographic factors best predict burglaries? We will utilize spatial and temporal analysis in our attempts to answer those research questions.

4 Literature Review

Modeling and predicting crime is an important task for applied statisticians and data science researchers. Common approaches include modeling spacial effects and temporal effects. The spatial neighborhoods of high crime can be used by policy makers to show where police should place emphasis (Amoako). Temporal analysis can be utilized for scheduling of law enforcement and other crime stopping measures (Groups, et al.). Our investigation seeks to capture the same usefulness as depicted in a study on the spatial analysis of Detroit census block groups and the granularity of daily temporal changes in crime. Our model will use the lowest time element of months and census blocks will be used to depict area. We are unable to achieve the same granularity of day by day temporal analysis due to limitations of the data. Spatial and temporal analysis is important to show otherwise difficult to model trends within a densely populated urban area like Chicago.

5 Methodology

Our general methodology was to explore the data, start fitting temporal and spatial models using different combinations of covariates, then draw conclusions from the created models. We knew going into the project that we wanted to utilize both spatial and temporal models to model the data. All of our models were fit using INLA. We began our exploratory analysis by looking at the data and creating a sources of variation chart. The data provided is count data burglaries of 552 census blocks in the Chicago area by month from 2010-2015. Our model will seek explain change in the number of burglaries in month i in census block j of Chicago. Our sources of variation diagram is Table 1.

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

With the sources of variation diagram in mind we plotted correlation charts, spatial graphs, and temporal graphs. We standardized wealth and young male population by total population per census blocks to allow for comparison between census blocks. When we built models we had one focusing on temporal effect, one on spatial effects, and another with both effects. We then compared the models using DIC to find the best model.

6 Data Exploration

During our data exploration we created a correlation graph of the data. Wealth and population are heavily correlated within the data, signifying that higher population means more wealth, or that wealth is not standardized across the census blocks. This sticks out because census blocks should have population sizes around the same. Figure 1 shows our correlation heatmap.

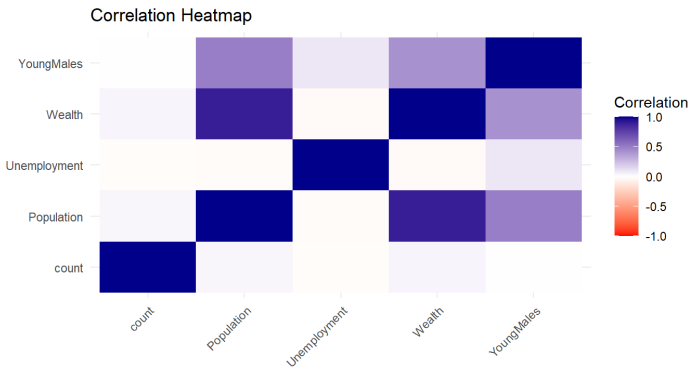

Figure 1: Correlation Heatmap

For our exploratory analysis we looked at the temporal and spatial trends within the data. 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 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. Figure 2 shows burgularies over time in Chicago.

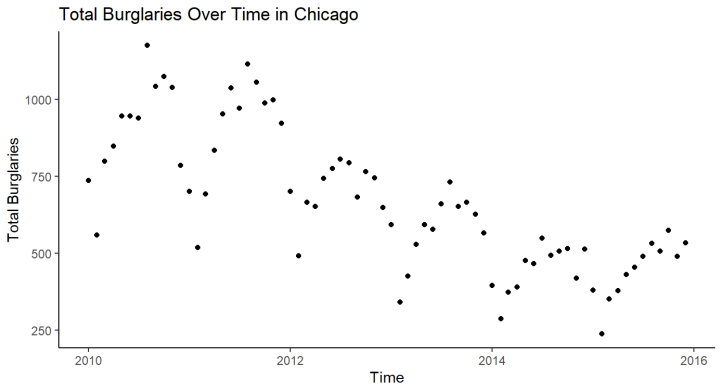

Figure 2: Aggregated Burglaries over Time

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, by using larger geographic areas such as counties, or by utilizing a BYM model. Figure 3 shows the total number of burgularies in every census block.

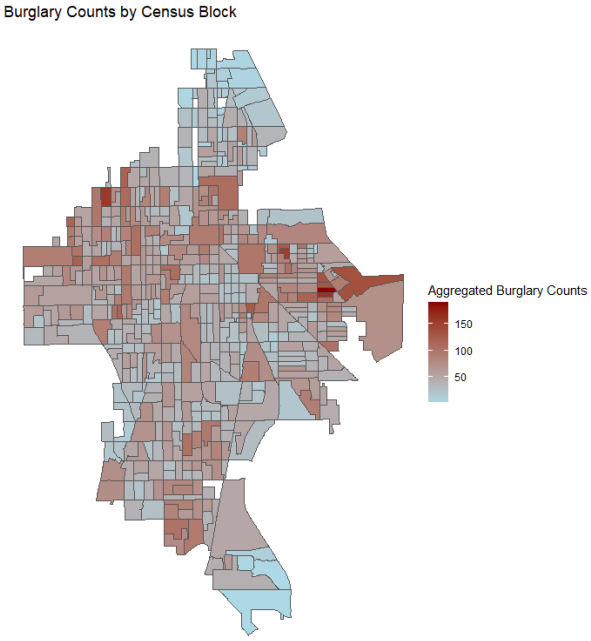

Figure 3: Aggregated Burglaries over Space

For data preparation, we transformed the data from wide format to long while combining subsequent data frames together joining them by block number. Our main data frame now contains block number, population, young male population, wealth metric, unemployment, and the count for every burglaries each month.

7 Model Building

Given our data exploration and research questions we believe the following models would be promising. For all of our models, we will use a random effect for individual blocks so that we can later generalize our findings for our covariates to blocks that are not included in our dataset. We will also use an offset for

population, since there is a wide range of populations in the census blocks. We will fit all of our models using integrated nested Laplace approximations.All of our models will be evaluated using DIC.In our first model we will try to establish a general trend over time. This model includes all economic variables with an offset for population and a random effect for location.

$$y_{ij} \sim \text{Po}(\lambda_{ij})$$
$$\log(\lambda_{ij}) = \eta_{ij}$$
$$\eta_{ij} = \beta_0 + \log(\text{Pop}_j) + \beta X + u_i$$
$$X = \begin{bmatrix} \text{Unemployment Proportion}_j \\ \text{Wealth Per Capita}_j \\ \text{Young Male Proportion}_j \\ \text{Months since Jan 2010}_i \\ \text{Temperature}_i \end{bmatrix}$$
$$u_i \sim N(0, \sigma_u^2) \text{ is a random component for Census Block}$$
$$\text{Prior}(\frac{1}{\sigma^2}) \sim \text{Loggamma}(1, 10^{-5})$$

Results for our first model are:

$$X = \begin{bmatrix} 1 \\ \text{Unemployment Proportion}_j \\ \text{Wealth Per Capita}_j \\ \text{Young Male Proportion}_j \\ \text{Months since Jan 2010}_i \\ \text{Temperature}_i \end{bmatrix} \hat{\beta} = \begin{bmatrix} (-7.021, -6.803) \\ (-0.201, 1.055) \\ (-23.455, -14.994) \\ (0.074, 2.498) \\ (-0.013, -0.013) \\ (0.007, 0.008) \end{bmatrix}$$
$$\frac{1}{\hat{\sigma}^2} = (3.08, 3.99)$$

The DIC for the first model is 111507. We see generally decreasing burglaries over time, with spikes at higher temperature times, this shows consistency with our temporal analysis and graph. Wealth per capita has a strong negative effect on burglaries. This suggests that neighborhoods that are wealthier are less likely to have burglaries. Young male population has a positive effect on observed burglaries. This suggests the inverse to wealth, more young males means more burglaries.

For model 2 we are focused on that spatial effect. We utilize the BYM model for location, all economic variables included in the model with an offset for population, and AR1 random variable for time.

$$y_{ij} \sim \text{Po}(\lambda_{ij})$$
$$\log(\lambda_{ij}) = \eta_{ij}$$
$$\eta_{ij} = \beta_0 + \log(\text{Pop}_j) + \beta X + \gamma_j + u_t$$
$$X = \begin{bmatrix} \text{Unemployment Proportion}_j \\ \text{Wealth Per Capita}_j \\ \text{Young Male Proportion}_j \end{bmatrix}$$
$$\gamma_j \text{is BYM Model}$$
$$u_t = \Phi u_{t-1} + \upsilon_t$$

The results for the second model are:

$$X = \begin{bmatrix} 1 \\ \text{Unemployment Proportion}_j \\ \text{Wealth Per Capita}_j \\ \text{Young Male Proportion}_j \end{bmatrix} \hat{\beta} = \begin{bmatrix} (-7.315, -6.675) \\ (-0.218, 1.048) \\ (-23.642, -15.186) \\ (0.004, 2.435) \end{bmatrix}$$
$$\text{Precision for Block (iid component)} = (3.024, 4.009)$$
$$\text{Precision for Block (spatial component)} = (18.389, 199.561)$$
$$\text{Precision for month number} = (3.806, 17.289)$$
$$\Phi = 0.866$$

The DIC for the second model is 110114. We see similar coefficients for economic variables as the first model, drawing the same conclusions about wealth and young male proportion. The high precision for the spatial components indicates a strong component for clustering between blocks. Showing that blocks are not independent of each other. The AR1 coefficient indicates very strong correlation between consecutive months.

For model 3 we attempt to combine the two previous models. We utilize the BYM component for location, a fixed effect for time, and we decided to omit unemployment.

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

$$X = \begin{bmatrix} \text{Wealth Per Capita}_j \\ \text{Young Male Proportion}_j \\ \text{Months since Jan 2010}_i \\ \text{Temperature}_i \end{bmatrix}$$

$\gamma_j$is BYM Model

The results for the third model are:

$$X = \begin{bmatrix} 1 \\ \text{Wealth Per Capita}_j \\ \text{Young Male Proportion}_j \\ \text{Months since Jan 2010}_i \\ \text{Temperature}_i \end{bmatrix} \hat{\beta} = \begin{bmatrix} (-6.994, -6.781) \\ (-23.294, -14.859) \\ (0.145, 2.558) \\ (-0.013, -0.012) \\ (0.007, 0.008) \end{bmatrix}$$
$$\text{Precision for Block (iid component)} = (3.05, 3.96)$$
$$\text{Precision for Block (spatial component)} = (118.86, 8023.20)$$

The DIC for the third model is 111507. We have similar interpretations of our coefficients as the previous models with the same conclusions for wealth and young male proportion. We added an unemployment proportion to this model and DIC was essentially unchanged (less than 1). This model also shows that adding fixed effects for time is not more efficient than using an AR1 model. The fixed effect shows the general decrease but not the undulations as the seasons change.

8 Model Selection

When looking at our three models we have built we will compare them utilizing DIC. Table 2 shows the DIC for all three models.

Model Name	**DIC**
Model 1	111507
Model 2	110114
Model 3	111507

Table 2: DIC Values for Three Models

We suggest that the stakeholders utilize model 2, because it has the lowest DIC out of the three models listed. Model 2 also allows to generalize across different times better, since we only need to know the previous observation, rather than a general trend thanks to AR1. Model 2 also allows for the spatial component between blocks to be modeled through BYM. However, the other models are still useful for answering our research questions.

9 Conclusion

In conclusion, we believe that Model 2 best answers our research questions due to it having the lowest DIC and explainability for the time component as shown through month and spatial component by using BYM. Through our three models we are able to answer our initial research questions. We can model temporal patterns by month and year utilizing fixed effects and temperature. We can also represent month to month similarity by using AR1. We are able to point out location patterns within the data by using BYM to show the clustering. We are also able to point out that wealth per capita and young male proportion are significant indicators of burglaries.

10 Discussion

Some limitations with our findings is the year range of the data and the fact that crime is a very complex subject. Conducting this study on 2010-2015 data in 2024 makes the extrapolation to our current year impossible. There are several things that can change almost a decade after the data was collected. Crime is a complex and dynamic issue that is hard to model based on one type of statistical model. Police policy and criminal actions change all of the time so it is difficult to adequately model that behavior almost a decade after.

Ethical considerations mainly occur in the analysis and deployment section of the Deon Driven Data checklist. There could be dataset bias due to law enforcement having a higher patrol rate in certain census blocks. If this model was deployed in 2024 it would violate several ethical considerations. Due to the data being collected from 2010-2015 there are several issues with it being used for analysis in today's society. Crime is a complex problem and using this model to single out a single census block could violate redress and harm those who live in the census block. This model should not be used for predictive of inferential analysis due to the time constraints to the data and lack of robust checks put inplace to avoid bias.

Future work could include expanding the model past burglaries and including more years in the study. Adding more violent crimes or white collar crimes would be interesting to see. It would also be interesting to apply this same framework to other metropolitan areas like the Bay Area or Miami.

11 Appendix A

All R code can be found on the following GitHub Page: https://github.com/THild24/MA478/tree/main

12 References

Amoako, E. A.: A spatial Analysis of Crime and Neighborhood Characteristics in Detroit Census Block Groups, Proc. Int. Cartogr. Assoc., 4, 5, https://doi.org/10.5194/ica-proc-4-5-2021, 2021.

Andresen, M.A., Malleson, N. Intra-week spatial-temporal patterns of crime. Crime Sci 4, 12 (2015). https://doi.org/10.1186/s40163-015-0024-7

Hyatt, Jacob CDT, assistance to author via previous work. I used the project proposal and final project slides as a reference when writing this paper. West Point, NY. 07MAY2024.