Chicago Burglaries: An Examination Using Generalized Linear Modeling

A.J. Pimentel

MA478: Generalized Linear Modeling, *United States Military Academy*

adrian.pimentel@westpoint.edu

Abstract

This paper examines temporal and socio-economic variables to underpin the effect on number of burglaries in 552 census block groups from 2010-2015. Understanding the effects of these variables can better aid lawmakers to formulate policies to limit the number of burglaries occurring in each census block. The total number of burglaries throughout this period show a general decrease overtime, while also exhibiting seasonal variability. Throughout this analysis, we utilize several modeling approaches ranging from a simple Poisson generalized linear model to a generalized linear mixed effects model to consider random effects. In this paper we neglect to consider any clustering or correlation among census blocks. We rather opt to utilize a random effect in our generalized linear mixed effects model to account for the unique variation among census blocks. To begin our analysis, we first transformed our variables, as well as manipulated our data to analyze burglaries on a monthly basis. We then did an initial exploratory variable analysis to determine multicollinearity. We began our modeling process, utilizing goodness of fit, predicted zeroes, and AIC to determine the best model. Additionally, variables were included on a manual selection basis. Results among our Poisson, zero-inflated Poisson, and generalized linear mixed effects model were relatively consistent. Each of our models found the percentage of young males, population, unemployment, year, and month (as a factor) to be a significant determinate in the log of the expected count of burglaries for a census block.

Key words: Policy formulation, Chicago Burglaries, Temporal Analysis, Socio-economic effect, Generalized linear modeling.

Introduction

Mary Pattilo, a prominent sociologist out of Northwestern University stated that crime and violence were one of the most pressing topics facing the city of Chicago (Patillo). Given this information, the city of Chicago must produce an appropriate plan of action to maintain the cities prestige. Once we analyze variables such as wealth, population, unemployment, and young males, we can begin to offer recommendations and adjustments to policy in these areas. For example, if we find that areas of low wealth contribute significantly to burglaries in an area, we can create some sort of policy to create wealth in an area, and therefore lower the overall burglaries in the area. We are particularly interested in the effect that young males have on the expected count of burglaries. As we do this analysis, we must be very cautious to correctly model the problem and ensure that correct statistical methods are utilized to draw conclusions. If this is done incorrectly, we could potentially recommend policies that will actually increase the number of burglaries. We must also be cautious of ethical implications and ensure that our recommendations are not discriminatory in any manner.

Literature Review

When considering our approach to the problem, we can draw from other data scientists and applied statistics researchers. Modeling spatial-temporal effects among other variables have been researched extensively over the past twenty years. This problem is certainly not unique to Chicago, as researchers have analyzed this in Portugal, Texas, Seoul, and Charlottesville, VA among many other locations.

While each analyst attempted to tackle similar problems, they each utilized different statistical methods to draw conclusions. Taveres and Costa analyzed eight different independent variables; young resident population, education, purchasing power, unemployment rate, and foreign population to name a few (Tavares & Costa). To model property crime, they utilized Geographically Weighted Poisson Regression models, and semi parametric GWPR models, where they found the GWPR model have the best performance (Tavares & Costa). Yongwan Chun, attempted to tackle vehicle burglaries in Plano, Texas using an extension of the generalized linear mixed model (Chun). Specifically, Chun examined the effectiveness of eigenvector spatial filtering (ESF), where he found this extension of the model to be quite effective when spatial autocorrelation is present (Chun). Nathaniel Garton and Jarad Niemi also tackled crime in Chicago from 2007 to 2016. To understand correlations in crime trends, Garton and Niemi took advantage of dynamic linear models, which are a generalization of ARIMA models (Garton and Niemi). Similar to our goals, Garton and Niemi used these models for inferential purposes rather than forecasting, which is the typical use for DLM’s (Garton and Niemi). To better understand the seasonal variation we find in burglaries, we can learn from Yendae Jung, Dohyeong Kim, and Alex Piquero’s approach to the challenge. The group aimed to analyze the association between temperature and assaults at 424 subdistricts in Seoul, South Korea using a generalized linear mixed-model approach (Jung et al.) Their overall analysis concluded that there was a positive and significant linear association between temperature and assaults, which is similar to what we see between time of year and total count of burglaries in each respective census block group (Jung et al.). Lastly, Xiaofeng Wang and Donald Brown out of the University of Virginia's Department of Systems and Information Engineering, examined the usefulness of generalized additive models for modeling criminal incidents in Charlottesville, VA (Wang and Brown). In this analysis, they found that their spatial temporal GAM outperformed previous spatial prediction models in predicting future criminal incidents (Wang and Brown).

While each of these approaches proved to be effective in modeling some aspect of the spatial-temporal effects on crime, each of the models have their own shortcomings. For example, Tavares and Costa found that their geographically weighted Poisson regression models were not effective at generalizing throughout Portugal, and the association of each variable varied significantly across municipalities (Tavares & Costa). Niemi and Garton found the problem with their DLM approach is the required assumption that the data is Gaussian (Garton & Niemi). This restricts the usage of DLM's to situations where crime counts are sufficiently large in any given time period, which isn't necessarily what we see in our Chicago burglary dataset (Garton & Niemi). Consistent across all models was the lack of generalizability for other populations. We found that as researchers attempted to broaden their location, the explanatory power of the model dropped significantly. This implies that there is some random variation within the data that can not be accounted for with fixed effects.

As we move forward in our analysis, we find that there aren't necessarily any right answers to modeling this problem. While some of the crimes measured were different (assaults for example), we find that they generally follow the same seasonal pattern throughout time that burglaries do. When examining the generalizability of the data, we find from our literature review that researchers in the past have had a hard time generalizing for their entire dataset. For example, Tavares and Costa were able to effectively employ their models at the municipal level but found their model to have poor explanatory power when attempting to generalize for all of Portugal. This recommends that there is some level of uniqueness at the block/municipal level that can not be captured through fixed effects. Our takeaway from this is that we may need to incorporate some random effects into our models later in the process. Having this background in the subject, we can then begin to conduct our own informed analysis on our data.

Methodology

Key Aspects and Assumptions

Prior to building out our models, there are a few key assumptions about the data that we are making. The first being is that police presence has a negligible effect on our response variable (burglary count for census block given month and year). Police presence likely has a great effect on someone's decision to break into a home or car, so if this was present in our dataset, it would very likely be a statistically significant variable. Unfortunately, we don't know this information, so we can not make any conclusions regarding the allocation of police forces as a result of our inferential analysis. There are also several other variables that we are unable to account for within our data. Variables such as education level, home presence, census block type (residential or commercial), and block gentrification are unknown variables. Each of these variables likely influence the observed burglary count.

We also must assume that our variables population, wealth, unemployment, and young males remain constant throughout the duration of our analysis. It is a known fact that people move, wealth levels change, unemployment levels change, and the number of young males in an area likely fluctuate greatly. Throughout this analysis, we assume that these explanatory variables remain constant throughout all 72 months of the observational period.

Data

The dataset examined in this inferential analysis consists of burglaries from 552 different census block groups during 2010-2015. We also have total population, unemployment percentage, a standardized level of wealth, and number of young males for each respective census block.

Initial Models

Throughout this analysis, we began with three base models of increasing complexity. First, a basic Poisson GLM was considered. Prior to going into our first model, we must introduce the format of our response and explanatory variables.

Response:

$$Y_{i,j,k} = burglaries\ in\ census\ block\ i, year\ j, and\ month\ k$$

Explanatory:

$$x_{1,i} = percent\ young\ males\ in\ population\ of\ census\ block\ i$$

$$x_{2,i} = population\ of\ census\ block\ i$$

$$x_{3,i} = unemployment\ in\ census\ block\ i$$

$$x_{4,j} = year\ j$$

$$x_{5,k} = \begin{cases} 1\ if\ month\ k \\ 0\ if\ otherwise \end{cases}$$

$x_{6,i}$ = wealth in census block i

With this model framework in mind, our base models our are outlined below:

$$Y_{i,j,k} \sim Po(\lambda_{i,j,k})$$

$$\log(\lambda_{i,j,k}) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,j} + \beta_5 x_{5,k} + \beta_6 x_{6,i}$$

Equation 1: Base Poisson Model

Our criteria for removing variables will be contingent upon collinearity, as well as variable significance, as we aim to have a parsimonious model. Once this parsimonious model is achieved, a ZIP model will be considered to account for any potential overdispersion. You'll notice that wealth was removed as a covariate in this model. Further information on this decision can be found in the modeling section. A ZIP model is an appropriate model to consider given the high number of observed zeroes.

$$Y_{i,j,k} \sim ZIP(\pi_{i,j,k}, \lambda_{i,j,k})$$

$$\text{logit}(\pi_{i,j,k}) = \gamma_0 + \gamma_1 x_{1,i} + \gamma_2 x_{2,i} + \gamma_3 x_{3,i} + \gamma_4 x_{4,j} + \gamma_5 x_{5,k}$$

$$\log(\lambda_{i,j,k}) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,j} + \beta_5 x_{5,k}$$

Equation 2: Base ZIP Model

Our ZIP model above attempts to better account for the high number of zeroes in our data. These zeroes may reflect an absence of reporting. Similar to the Poisson model, we'll begin with our saturated models, and remove insignificant variables.

$$Y_{i,j,k} \sim Po(\lambda_{i,j,k})$$

$$\log(\lambda_{i,j,k}) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,j} + \gamma_i + \delta_k$$

$$\gamma_i \sim N(0, {\sigma^2}_{\gamma_i})$$

$$\delta_k \sim N(0, {\sigma^2}_{\delta_k})$$

Equation 3 Generalized linear mixed effects model

For our final model, a generalized linear mixed effects model will be considered to account for any random effects. For our model, we include the random effects of block. We also will include month due to the non-linear behavior it exhibits. If any random effects have a variance of zero, we will remove the random effect as it likely harnesses better explanatory power as a fixed effect.

Exploratory Analysis

The first step in our exploratory data analysis is to construct plots of our response variable to see if there is any particular behavior. Following this, we will then plot the distribution of some of our key explanatory variables. Next, we can create a correlation plot to see if there are any problems with multicollinearity among our explanatory variables. Lastly, if there are any significant correlations, we can examine a scatter plot of the two variables to determine the nature of the interaction.

Data Preparation

Our data preparation is relatively straightforward. We will standardize any variables necessary, such as unemployment. The reason for standardizing some of the variables is because since unemployment is in decimal form, it likely will have a negligible effect on the overall count. However, when we standardize this variable, we could potentially find it to be more significant. Standardization also allows for a less computationally expensive model. The same holds true for the year variable. Since 2010 is such a large number, scaling it back to 0-5 will make year have a much more statistically significant effect. Lastly, we will convert our data from wide to long format so we can analyze each census block at the month and year level.

Model Building Criteria

As mentioned above previously, we will begin with the saturated models. Variables will only be removed if they are insignificant or collinear. To determine which model performs the best, a Chi-squared test will be conducted for nested models, as well as a comparison of AIC for comparison between different types od models. An examination of predicted vs actual zeroes will also be considered when selecting a model. When determining what random effects to include in our generalized linear mixed effects model, we'll consider the random effect of census block based on the research from our literature review, as well as the month given its non-linear behavior.

Experimentation and Results

Data Exploration

Prior to exploring our data, we first need to convert our data from wide to long format. Conducting this operation allows for us to analyze our data at the monthly and yearly level. The only use for our wide set is to plot the total number of burglaries overtime by month and year. In Figure 1, we see that the total number of burglaries follows a general decrease overtime, but increases in warmer months, and declines in colder months.

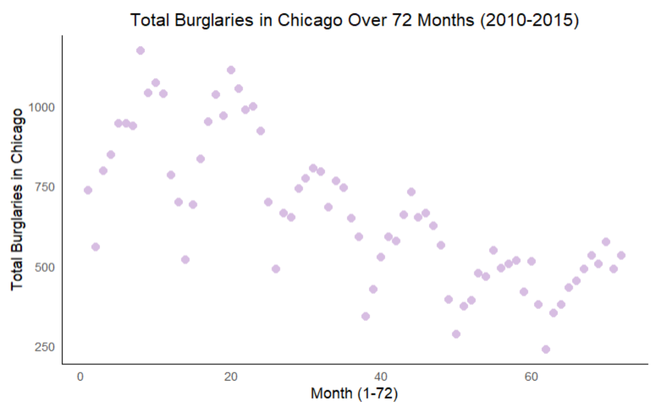


Figure 1: Total burglaries over time

In figure 2 of Appendix A, we examine the distribution of burglaries by census block in a month. Of note, there are over 15,000 occurrences of zeroes, so it is important our model can capture this aspect. After plotting another variable of interest, young males, we find that there are numerous blocks that have zero young males in the block. This histogram can be found in figure 3 in Appendix A. To account for this, we will impute the variable to be reflective of the median number of young males. The key assumption we are making here is that what we believe to be commercial census blocks still contain young males. This is justifiable considering that there are likely young males that are present in these commercial populations and are likely there for work. Figure 4 shows a scatter plot of how these zeroes are impacting the linear pattern exhibited in the relationship between young males and total burglaries. Lastly, we'll examine correlation among the variables. Figure 5 shows a strong correlation between wealth and population, with a level of .9. Figure 6 in Appendix A explores this relationship. From a quick glance, we find that there is a strong positive linear association between the two variables. This likely implies that wealth is a measure of total wealth rather than per capita. To account for this, we'll divide wealth by population. Figure 7 represents this change. You'll notice that even after attempting to account for this change, there is still a strong pattern in the data. To account for this, we'll only choose to include the more significant variable when conducting our Wald-test on the variables. Now that we have visualized our data, and prepped our data, we can begin building our models.

Model Building

We will begin with our simplest model, a Poisson GLM. This approach is appropriate given that our response variable is count data. As defined above, our saturated model, equation 1, is considered first.

$$Y_{i,j,k} \sim Po(\lambda_{i,j,k})$$

$$\log(\lambda_{i,j,k}) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,j} + \beta_5 x_{5,k} + \beta_6 x_{6,i}$$

Equation 1: Base Poisson Model

The results of this model can be found in Table 1 of Appendix B. With the exception of winter months, the expected count appears to increase as our explanatory variables increase, with population harnesses the greatest explanatory power. Recognizing that wealth and population are strongly correlated, we must determine which covariate to remove in our next model. Given that population has almost three times the z-value of wealth, we will continue our modeling process with just the population variable rather than wealth. This revised equation is outlined below:

$$Y_{i,j,k} \sim Po(\lambda_{i,j,k})$$

$$\log(\lambda_{i,j,k}) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,j} + \beta_5 x_{5,k} + \beta_6 x_{6,i}$$

Equation 4: Revised Poisson GLM

This revised models outputs can be found in table 2 of Appendix B. The results of this are generally the same, however, we find that as percentage of young males increases, the expected burglary count decreases.

Now that we have our model, we will conduct a goodness of fit test. Our test yields a value of 0, meaning we have evidence for a lack of fit. An assumption of the Poisson model is that the mean and variance ratio are equal. To estimate our dispersion parameter for this model, we'll divide the deviance of our model by the residual deviance. If our data is not over dispersed, this should yield a value of 1. This method yields an estimate of 1.54, indicating that our model is over dispersed. A potential solution to solving this issue of overdispersion is to consider a ZIP model. Our initial ZIP model is outlined below:

$$Y_{i,j,k} \sim ZIP(\pi_{i,j,k}, \lambda_{i,j,k})$$

$$\text{logit}(\pi_{i,j,k}) = \gamma_0 + \gamma_1 x_{1,i} + \gamma_2 x_{2,i} + \gamma_3 x_{3,i} + \gamma_4 x_{4,j} + \gamma_5 x_{5,k}$$

$$\log(\lambda_{i,j,k}) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,j} + \beta_5 x_{5,k}$$

Equation 2: Saturated ZIP model

Recall we removed wealth as it proved to be less significant than population, and it exhibited a non-linear relation when adjusted for wealth per capita. The output of this model is found in Table 3 and 4 in Appendix B.

With this model we find that month seems to be an insignificant variable in predicting zeros. We also find that the percentage of young males does not appear to be significant in our Poisson with log link portion of our ZIP model. To achieve the more parsimonious model, we'll remove these two variables, and then compare the two using a likelihood ratio test. Our adjusted ZIP model is outlined below:

$$Y_{i,j,k} \sim ZIP(\pi_{i,j,k}, \lambda_{i,j,k})$$

$$\text{logit}(\pi_{i,j,k}) = \gamma_0 + \gamma_1 x_{1,i} + \gamma_2 x_{2,i} + \gamma_3 x_{3,i} + \gamma_4 x_{4,j}$$

$$\log(\lambda_{i,j,k}) = \beta_0 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,j} + \beta_5 x_{5,k}$$

Equation 5: Adjusted ZIP Model

The summary output of this model is in Appendix B, Tables 5 and 6. With this new model, we find that the covariates follow the same general interpretation in the context of our problem as the previous does. As we compare these two models, we find the p-value from our likelihood ratio test to be .0085, insinuating the first, more complex model to be preferred.

A unique aspect to this problem is that each census block exhibits its own unique behavior. Our literature review above confirms that there is an element of randomness to each census block. To account for this, we'll employ a generalized linear mixed effects model. We'll first consider block, as well as a random effect of month. Our year variable will be treated as a fixed effect considering the strong downward slope when aggregating total burglaries in a census block over years.

$$Y_{i,j,k} \sim Po(\lambda_{i,j,k})$$

$$\log(\lambda_{i,j,k}) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,j} + \gamma_i + \delta_k$$

$$\gamma_i \sim N(0, {\sigma^2}_{\gamma_i})$$

$$\delta_k \sim N(0, {\sigma^2}_{\delta_k})$$

Equation 4: Initial generalized linear mixed effects model

The summary output of this model is found in table 7 in Appendix B. From our results in this table, we find that the random effects from both block and month have a non-zero variance, meaning the random effect is significant.

Model Evaluation & Selection

To compare these models each of our selected models, we'll use AIC as a metric. We'll also simulate the number of predicted zero's vs actual zeros. Our first Poisson regression model has an AIC value of 119,718.9. Figure 8 in Appendix C displays a plot of fitted vs. actual values. We see that our fitted values do not reflect the reality of actual burglaries well. Our model fails to predict any counts higher than 3 burglaries in a given month. This model simulates 13,233 zeros compared to the actual amount of 16,044. Our second selected model, the ZIP model, yields an AIC of 117,396.7. Figure 9 in Appendix C displays the actual vs fitted values. We find that this model still does not reflect the higher burglary counts but does a much better job than the selected Poisson. The max fitted value for the ZIP model is 6 compared to the Poisson model's 3. This model predicts 13322 zeros compared to the actual amount of 16,044. Our final model, the generalized linear mixed model, has an AIC of 111613.9, and simulates 14,816 zeros. In Appendix C, Figure 10, we find that our generalized linear mixed effects model does a much better job of accounting for higher burglary counts than our previous models. However, our highest fitted value is 7, which is much lower than our actual observed highest value of 16 burglaries in a month. Considering that the Poisson model failed the goodness of fit test, and the ZIP did not provide an accurate estimate of our total number of zero's we can confidently select our generalized linear mixed effects model, as well as the lowest AIC.

Interpretation

Generally, all of the models had the same general pattern for each of the coefficients. The pattern we saw was that as standardized measure of percent of young males increases, the log of the expected count of burglaries decreases. We also find that as the standardized measure of population increases, the log of the expected count of burglaries increases. All of our model's state that as unemployment increases, the log of the expected count of burglaries will also increase. Lastly, we find that our year tends to have a strong negative linear effect on expected burglaries, and as we enter the warmer months, burglaries also tend to go up.

Specific to our selected generalized linear mixed model, we find congruence in this general pattern. We find that as the standardized measure of the percentage of young males increases by 1, the expected count of burglaries decreases by .96. Our GLMM states that as our standardized measure of population increases by 1, the expected burglary count will increase by 1.22. As our standardized measure of unemployment increases by 1, the expected burglary count for the census block will increase by 1.01. Lastly, we find that as the year increases by 1, the expected count of burglaries will decrease by .85 holding all else equal in the census block.

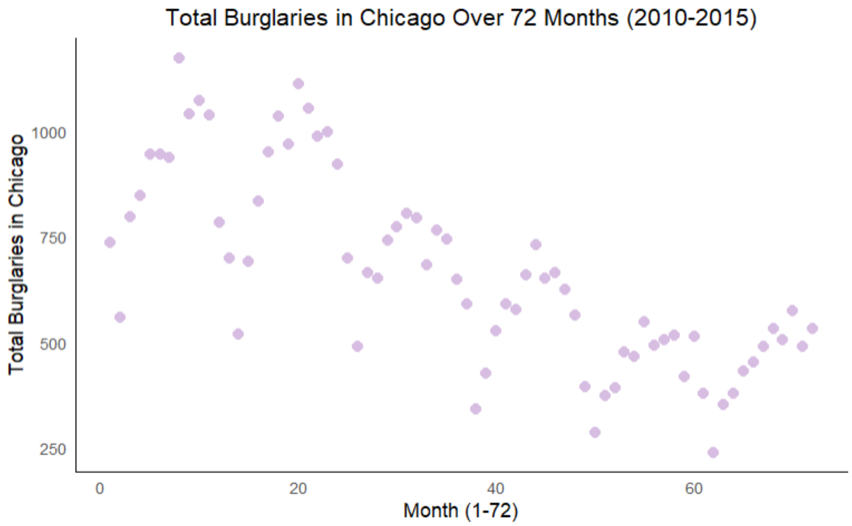
Ethical Considerations

While our results are interesting, there are numerous points at which ethics must be considered. Beginning with data collection, the biggest concern is that human subject's consent to whatever experiment they're being exposed to, and that they consent to having their data analyzed. In this inferential analysis, we do not have any human test subjects, but rather the census block. We do not have any information on the predominant race, gender, or religion in each census block, so it is relatively safe to conduct the analysis. Throughout our data analysis, there was no attempt made to hide any data that does not fit into our narrative. Throughout our modeling process, it is important to ensure that we are not utilizing any variables that are discriminatory in any manner. We have removed wealth from consideration, so the only variable to be somewhat weary of is unemployment. It would be unjust to penalize areas where most of the population is already unemployed by enacting policy that worsens the quality of life in the area simply because they cannot find a job.

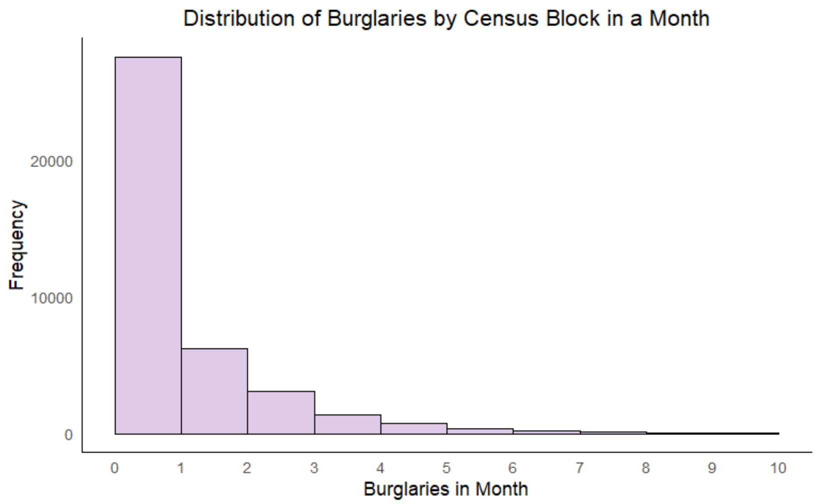
Discussion & Conclusion

In conclusion, our final model serves as an appropriate starting point for future work. When considering the limitations of the selected model, we find that there are numerous rooms for improvement in future work. For example, perhaps one month is more previously related to the previous month's burglaries. An autoregressive correlation structure would be appropriate to consider if we believe this to be the case. It would also be beneficial to introduce different imputation techniques for the 0 values of young males. For example, a nearest neighbor imputation would likely provide better results than simply imputing with the mean. We could also investigate additional approximation methods for our GLMM, such as Gauss-Hermite Quadrature to consider grouped/clustered outcome variables. In our exploration of the effect of young males, we find that as the standardized value for the percentage of young males increases by 1, the expected count decreases by .96 burglaries. While we certain will not be able to recommend a policy to end burglaries in Chicago, this model does provide some explanatory power to better understand the issue.

Appendix A



1: Total burglaries over time



2: Distribution of burglaries by census block in a month

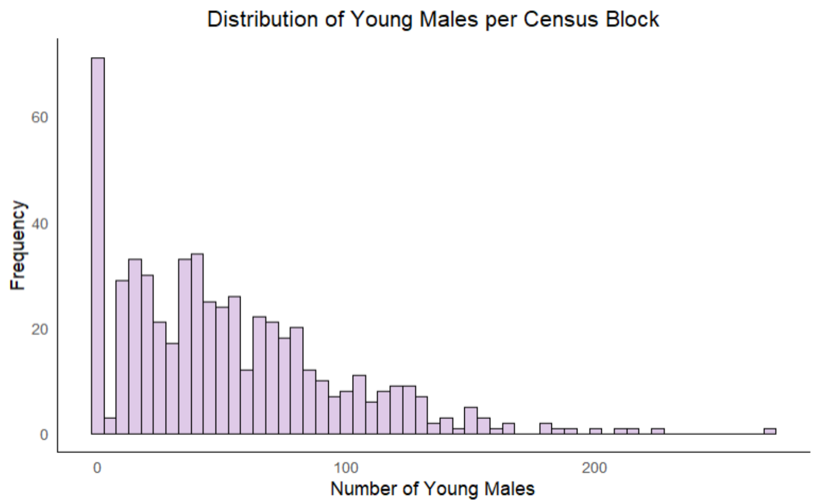


Figure 3: Distribution of young males per census block

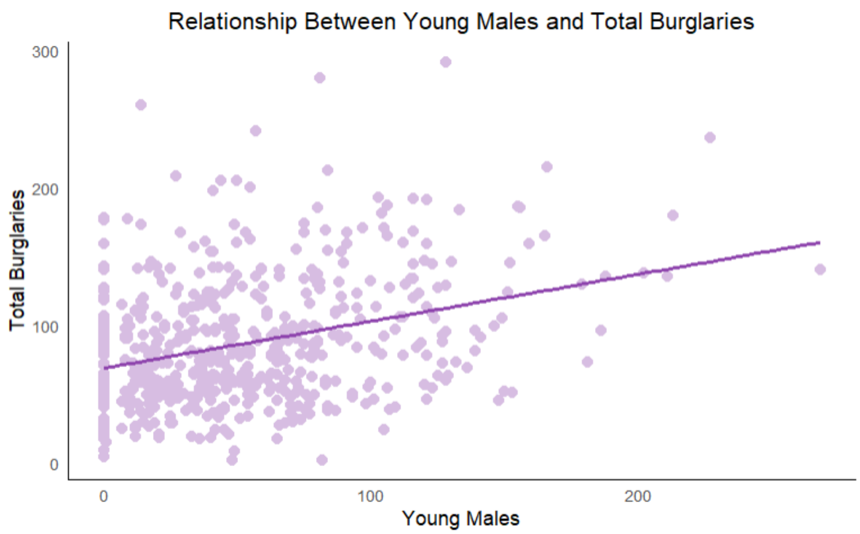


Figure 4: Scatterplot of Relationship between young males and total burglaries.

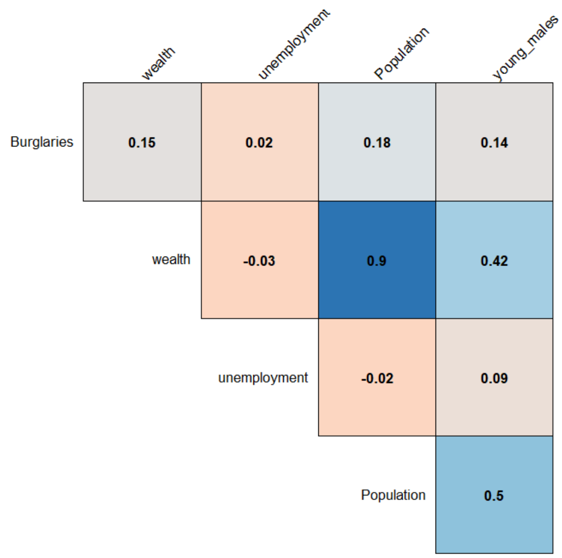


Figure 5: Correlation amongst explanatory variables.

Figure 6: Scatter plot of population and wealth

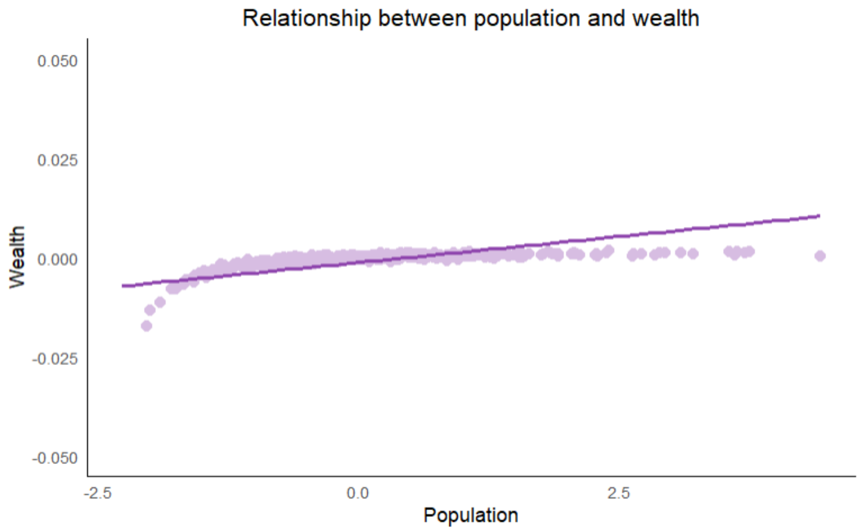


Figure 7: Relationship between population and wealth

Appendix B

Poisson GLM with Log Link

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.446887	0.017994	24.836	< 2e-16	***
perc_ym	0.366286	0.028462	12.869	< 2e-16	***
pop	0.174033	0.004437	39.224	< 2e-16	***
unemp	0.030442	0.004692	6.487	8.73e-11	***
year	-0.165125	0.002745	-60.164	< 2e-16	***
month02	-0.362852	0.026377	-13.756	< 2e-16	***
month03	-0.057578	0.024245	-2.375	0.0176	*
month04	0.035613	0.023684	1.504	0.1327	
month05	0.166834	0.022958	7.267	3.67e-13	***
month06	0.194710	0.022813	8.535	< 2e-16	***
month07	0.231388	0.022626	10.226	< 2e-16	***
month08	0.322882	0.022185	14.554	< 2e-16	***
month09	0.238160	0.022593	10.542	< 2e-16	***
month10	0.268734	0.022442	11.974	< 2e-16	***
month11	0.208709	0.022741	9.178	< 2e-16	***
month12	0.124390	0.023185	5.365	8.09e-08	***
wealth	37.688625	2.610267	14.439	< 2e-16	***

Table 1: Saturated Poisson GLM output

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.407573	0.017812	22.882	< 2e-16	***
perc_ym	-0.030314	0.007498	-4.043	5.27e-05	***
pop	0.194497	0.004162	46.734	< 2e-16	***
unemp	0.033803	0.004620	7.316	2.56e-13	***
year	-0.165125	0.002745	-60.164	< 2e-16	***
month02	-0.362852	0.026377	-13.756	< 2e-16	***
month03	-0.057578	0.024245	-2.375	0.0176	*
month04	0.035613	0.023684	1.504	0.1327	
month05	0.166834	0.022958	7.267	3.67e-13	***
month06	0.194710	0.022813	8.535	< 2e-16	***
month07	0.231388	0.022626	10.226	< 2e-16	***
month08	0.322882	0.022185	14.554	< 2e-16	***
month09	0.238160	0.022593	10.542	< 2e-16	***
month10	0.268734	0.022442	11.974	< 2e-16	***
month11	0.208709	0.022741	9.178	< 2e-16	***
month12	0.124390	0.023185	5.365	8.09e-08	***

Table 2: Poisson GLM 2 summary

Count model coefficients (poisson with log link):

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.5840144	0.0238805	24.456	< 2e-16	***
perc_ym	-0.0002198	0.0150732	-0.015	0.9884	
pop	0.1558403	0.0055792	27.932	< 2e-16	***
unemp	0.0500028	0.0063174	7.915	2.47e-15	***
year	-0.1325847	0.0037183	-35.658	< 2e-16	***
month02	-0.3087312	0.0378512	-8.156	3.45e-16	***
month03	-0.0635780	0.0328631	-1.935	0.0530	.
month04	0.0306754	0.0315929	0.971	0.3316	
month05	0.1599251	0.0300712	5.318	1.05e-07	***

```
month06       0.1639068  0.0300429   5.456 4.88e-08 ***
month07       0.1905366  0.0297156   6.412 1.44e-10 ***
month08       0.2831339  0.0287676   9.842  < 2e-16 ***
month09       0.2436313  0.0294309   8.278  < 2e-16 ***
month10       0.2273835  0.0292611   7.771 7.80e-15 ***
month11       0.1958679  0.0297313   6.588 4.46e-11 ***
month12       0.0923005  0.0308862   2.988   0.0028 **
```

Table 3: Output of Poisson with Log Link Coefficients for ZIP Model

```
Zero-inflation model coefficients (binomial with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.68345    0.09259 -18.182  < 2e-16 ***
perc_ym      0.07178    0.02491   2.882 0.003957 **
pop         -0.23944    0.02672  -8.961  < 2e-16 ***
unemp        0.08083    0.02307   3.504 0.000458 ***
year         0.16333    0.01395  11.712  < 2e-16 ***
month02      0.25066    0.13146   1.907 0.056548 .
month03     -0.01706    0.12506  -0.136 0.891482
month04     -0.01095    0.11940  -0.092 0.926954
month05     -0.01153    0.11334  -0.102 0.918998
month06     -0.16537    0.11800  -1.401 0.161100
month07     -0.22727    0.11855  -1.917 0.055226 .
month08     -0.18764    0.11325  -1.657 0.097566 .
month09      0.02775    0.10958   0.253 0.800047
month10     -0.21217    0.11602  -1.829 0.067427 .
month11     -0.04800    0.11309  -0.424 0.671240
month12     -0.17841    0.12246  -1.457 0.145148
```

Table 4: Summary of Binomial with Logit Link in ZIP Model

```
Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.572616   0.020094  28.497  < 2e-16 ***
pop          0.155924   0.005529  28.203  < 2e-16 ***
unemp        0.049926   0.006298   7.928 2.23e-15 ***
year        -0.133628   0.003692 -36.199  < 2e-16 ***
month02     -0.360397   0.028655 -12.577  < 2e-16 ***
month03     -0.061183   0.026494  -2.309   0.0209 *
month04      0.032723   0.025926   1.262   0.2069
month05      0.163156   0.025151   6.487 8.76e-11 ***
month06      0.186996   0.024991   7.483 7.29e-14 ***
month07      0.220559   0.024776   8.902  < 2e-16 ***
month08      0.307116   0.024262  12.659  < 2e-16 ***
month09      0.242168   0.024813   9.760  < 2e-16 ***
month10      0.254846   0.024548  10.382  < 2e-16 ***
month11      0.204079   0.024921   8.189 2.63e-16 ***
month12      0.117858   0.025401   4.640 3.49e-06 ***
```

Table 5: Output summary of Revised Poisson with log link ZIP Component

```
Zero-inflation model coefficients (binomial with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.76326    0.04078 -43.242  < 2e-16 ***
perc_ym      0.07320    0.01601   4.574 4.79e-06 ***
pop         -0.24165    0.02701  -8.948  < 2e-16 ***
unemp        0.08202    0.02326   3.526 0.000421 ***
year         0.15970    0.01397  11.430  < 2e-16 ***
```

Table 6: Output summary of revised binomial with logit link ZIP component

```
Random effects:
 Groups Name        Variance Std.Dev.
 block  (Intercept) 0.24617  0.4962
 month  (Intercept) 0.03254  0.1804

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.408787   0.056717   7.207  5.7e-13 ***
perc_ym     -0.042076   0.023470  -1.793    0.073 .
pop          0.203685   0.022082   9.224  < 2e-16 ***
unemp        0.018365   0.021936   0.837    0.402
year        -0.165125   0.002737 -60.328  < 2e-16 ***
```

Table 7: Summary Output of generalized linear effects model.

Appendix C

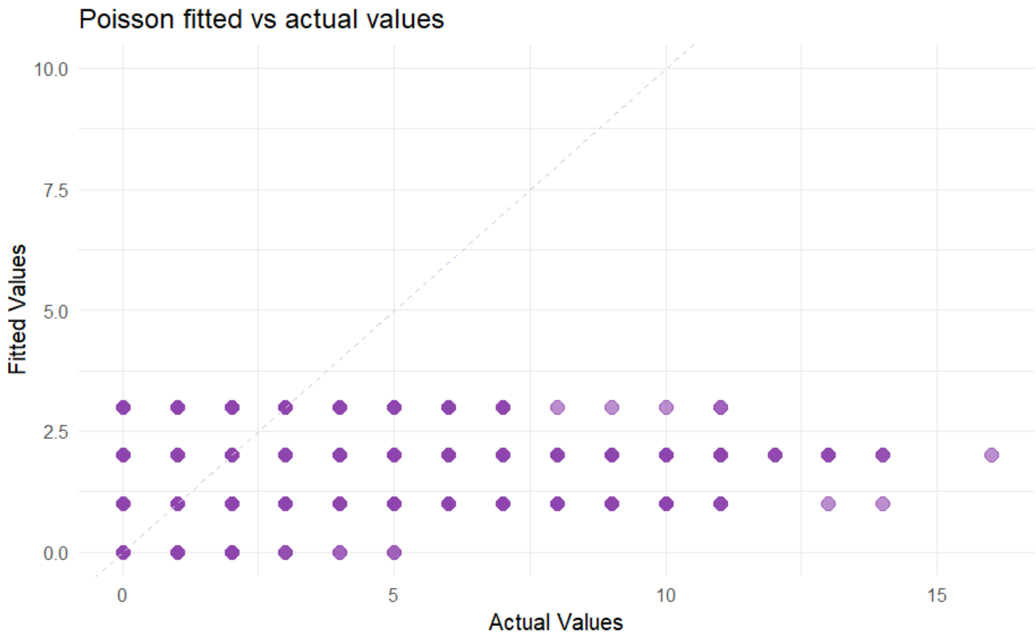


Figure 8: Fitted vs actual values for Poisson Model

Figure 9: ZIP model fitted vs. Actual values.

Figure 10: Actual vs. Fitted values.

Works Cited

Chun, Y. (2014), Analyzing Space–Time Crime Incidents. Geogr Anal, 46: 165-184. https://doi.org/10.1111/gean.12034

Garton N, Niemi J (2019) Multivariate temporal modeling of crime with dynamic linear models. PLoS ONE 14(7): e0218375. https://doi.org/10.1371/journal.pone.0218375

Jung, Y., Kim, D., & Piquero, A. R. (2020). Spatiotemporal Association Between Temperature and Assaults: A Generalized Linear Mixed-Model Approach. Crime & Delinquency, 66(2), 277-302. https://doi.org/10.1177/0011128719834555

Patillo, Mary. “Crime in Chicago: What Does the Research Tell Us?: Institute for Policy Research - Northwestern University.” *Crime in Chicago: What Does the Research Tell Us?*, 2018, www.ipr.northwestern.edu/news/2018/crime-in-chicago-research.html.

Tavares, J.P.; Costa, A.C. Spatial Modeling and Analysis of the Determinants of Property Crime in Portugal. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 731. https://doi.org/10.3390/ijgi10110731

Wong, Josh. CDT ’24. Co. I-3. *Verbal Assistance given to author*. CDT Wong explained to me how to structure the categorical variable within my modeling process. CDT Wong also assisted with the interpretation of my models output, as some of the variables were transformed. 7 May, 2024. West Point, NY.

X. Wang and D. E. Brown, "The spatio-temporal generalized additive model for criminal incidents," Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics, Beijing, China, 2011, pp. 42-47, doi: 10.1109/ISI.2011.5984048. keywords: {Predictive models;Computational modeling;Additives;Educational institutions;Hospitals;Roads},