Examining Crime Hotspots in Chicago Using Bayesian Statistics

Abstract

Chicago currently leads the United States with the greatest number of homicides and violent crimes in recent years. Using police data from the City of Chicago's Data Portal, we examined crime hot spots in Chicago and whether crime rates differ by geographic and demographic information. In general, we found that crime rate in Chicago has decreased between 2010 and 2015, though the rates differed between violent and non-violent crimes. Change in crime rate also varied geographically. We found that for areas with lower white populations, crime decreased as income rose. For areas with larger non-white populations, crime rate increased as income increased

1 Introduction

Chicago currently leads the United States with the greatest number of homicides and violent crimes in recent years. In 2016, the number of homicides in Chicago increased 58% from the year before (Ford, 2017). Using police data from the City of Chicago's Data Portal, we examined crime hot spots in Chicago and whether crime rates differ by geographic and demographic information. In this report, we defined hot spots as zipcodes with greater increases, or smaller decreases, in crime rates over time relative to other zipcodes in Chicago. In addition, we examined hot spots for both violent and non-violent crimes. Understanding crime hot spots can prove advantageous to law enforcement as they can better understand crime trends and create crime management strategies accordingly (Law, et al. 2014).

A common approach to defining crime hot spots uses crime density. Thus, hot spots by this definition are areas with high crime rates that are also surrounded by other high-crime areas for one time period. We were interested in finding areas where crimes increased relative to Chicago's general trend over time. Law, et al. (2014) identified hot spots by our definition of interest in the Greater Toronto Area using a Bayesian spatiotemporal modeling approach. Bayesian statistics allowed them to examine how hot spots change over time, especially in the presence of the small number problem. The small number problem occurs in areas with few crime count or small population, thus chance variation in crime over time might create dramatic change in crime density (Law, et al. 2014). For this reason, we applied similar Bayesian methods to examine how crime concentration in Chicago changed year over year.

Although Chicago is widely known for its high level of violent crime, certain areas of Chicago are prone to more crime than others. In Figure 1 below, the concentration of crime in Chicago is particularly high in areas highlighted in red. The concentration of crime appears to shift slightly from 2010 to 2016, with greater crime rates in northeast Chicago in 2016 compared to in 2010. In our analysis, since we defined hot spots as areas with a higher increase (or lower decrease) in crime rate compared to the average change over time, the areas that turn redder over time can be identified as hot spots.

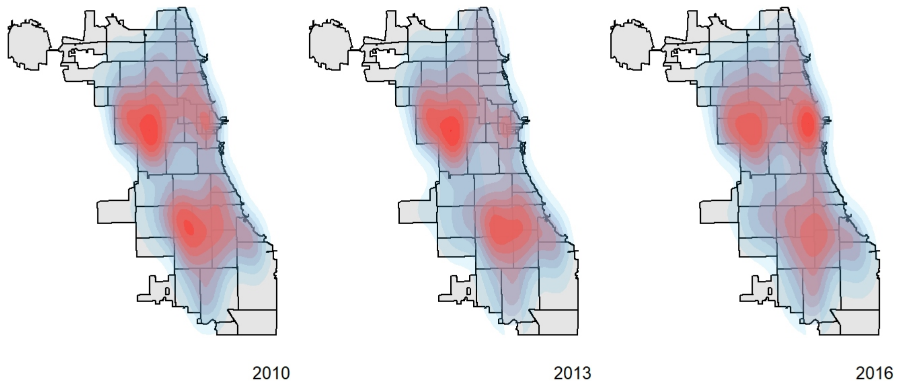

Figure 1: Concentration of crimes in Chicago in 2010, 2013, and 2016 (from left to right)

Figure 2, shown below, highlights the significant demographic differences across the zipcodes in Chicago. While per capita income and median rent appear positively correlated with both the Asian and White populations in Chicago, they appear negatively correlated with the Hispanic and Black populations. Since crime rates might differ across different ethnic and income groups, we take these demographics into consideration when predicting hot spots in Chicago.

2 Data and Methodology

To complete our analyses, we utilized data from the City of Chicago's Data Portal, in which each row represents a crime observation in Chicago. In addition to our crime data, we used the 2010 Census data for the percentage of white populations, per capita income, and total population for each zipcode in Chicago. We assume that demographic data remain constant from 2010 to 2017. Due to limitations in our computing capacity, we filtered for crimes that occurred between 2010-2017 and used a subset of 10,000 observations for our analysis. We then transformed the data to obtain the number of crime

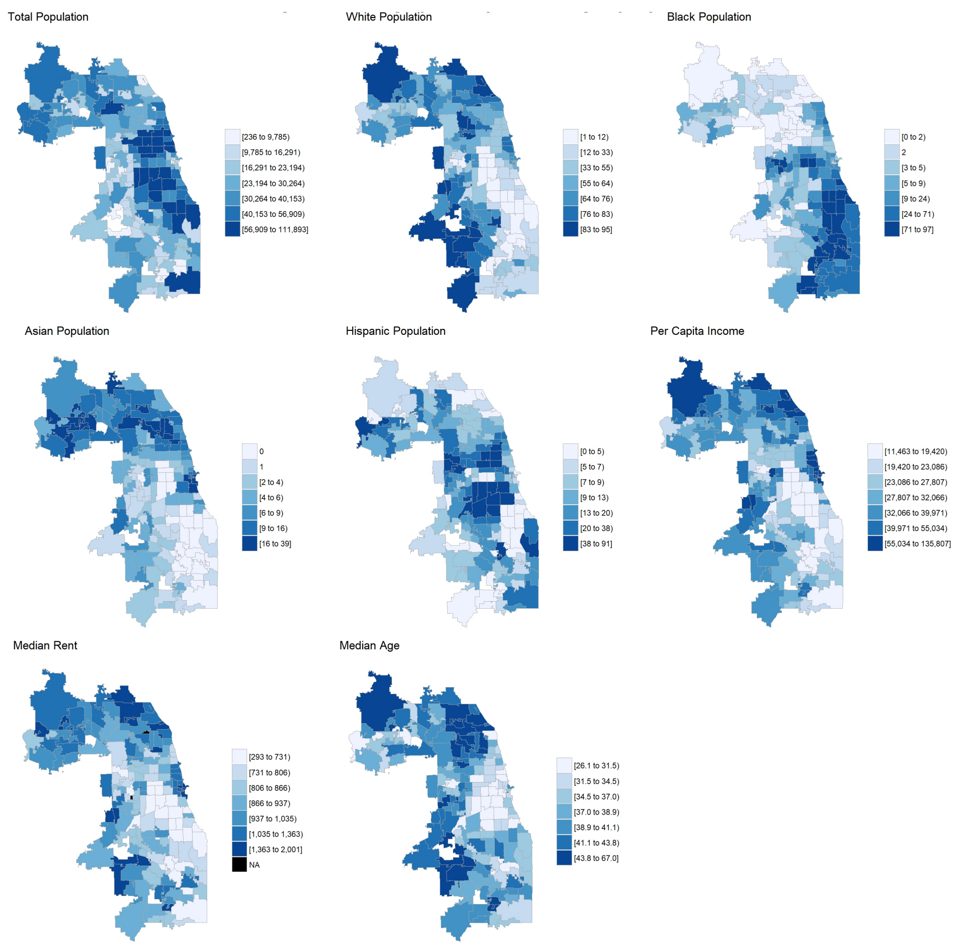

Figure 2: Maps of demographics by zipcodes of Chicago (from left to right, top to bottom): total population, White population, Black population, Asian population, Hispanic population, per capita income, median rent, and median age

counts for each zipcodes for each year from 2010 to 2017. We standardized the values for year and income to help the Markov Chains converge. We then focused on the following variables for our analysis: time, type of crime, income, percentage of white population and location.

To derive our models, we used the Gibbs sampler. We then checked for convergence of our Markov Chains by using effective sample sizes and a combination of visual indicators, such as running mean plots, trace plots of our parameters, and the Gelman-Rubin diagnostic statistics. These tools can be accessed using MacBayes package (at https://github.com/ajohns24/MacBayes) and diagnostic functions in the package coda. Upon checking the diagnostics for convergence, we observed autocorrelation between some of our parameters and reparametrize our models with hierarchical centering and orthogonalization of correlated predictors (Browne 2004; Browne et. al. 2009). We also found that standardizing our variables help the Markov Chains converge more quickly. We ran all of our models for 30,000 iterations for stabilization and subsequently calculate the posterior means. Finally, we checked all of the 95% credible intervals of our posterior means for statistical significance.

3 Analysis and Results

3.1 Time Trend

Figure 3 plots the relationship between the log of the crime rate and year for zipcodes in Chicago. It shows a slight downward trend in the log of crime rates as the years increase. Thus, we thought it was plausible to model overall log crime rates as a linear function of time. Although Figure 3 displays a lot of noise, we believed it was reasonable to see that each zipcode has a different slope and intercept.

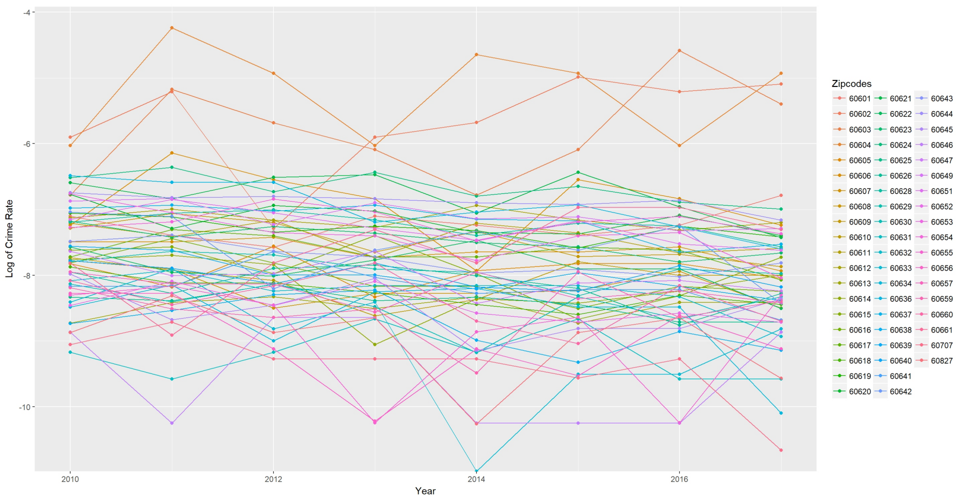

Figure 3: Crime rate over time, from 2010 to 2015, by zipcodes

In our first model, we predicted crime hot spots in Chicago using a simple model that incorporates a time trend. From our observations of Figure 3, we assumed in our model that log crime rate for each zipcode is a linear function of time. Here, y_{ij} is the observed crime count in zipcode i in year j and is modeled with a Binomial distribution. The parameters are p_{ij}, which is the inherent crime rate, and n_{ij}, which is the total population in zipcode i and year j. We model y_{ij} with a Binomial distribution because a Binomial distribution is typically used to model the count of an event given a probability between 0 and 1. In addition, we assumed zipcodes to have different time trends, denoted as δ_i for zipcode i, that are normally distributed around a mean crime trend of Chicago, denoted as δ_0, with precision τ_1. We also included random effect parameters, β_i, that try to account for variance in zipcodes' crime rates not explained by time trend. To prevent the random effects β_i to be drastically different between zipcodes, we gave them a normal prior distribution with mean β_0 and precision τ_0.

$$y_{ij} \mid p_{ij}, \beta_0, \beta_i, \delta_0, \delta_i, \tau_0, \tau_1 \sim Bin(p_{ij}, n_{ij})$$

$$log(\frac{p_{ij}}{1-p_{ij}}) \mid \beta_0, \beta_i, \delta_0, \delta_i, \tau_0, \tau_1 = \beta_i + \delta_i * time_j$$

$$\beta_i \mid \beta_0, \tau_0 \sim N(\beta_0, \tau_0^{-1})$$

$$\delta_i \mid \delta_0, \tau_1 \sim N(\delta_0, \tau_1^{-1})$$

$$\beta_0, \delta_0 \sim N(0, 1000^2)$$

$$\tau_0, \tau_1 \sim Gamma(0.5, 0.0005)$$

where $i = \{1, ..., 58\}$, $j = \{1, ..., 5\}$

The precisions, τ_0 and τ_1, are modeled with a Gamma distribution whose mean equals 0.00025. This is equivalent to a mean standard deviation of 63 for β_i and δ_i's normal distributions. This large variance indicates that we gave vague prior to β_i and δ_i. The priors for β_0 and δ_0 are also vague with a mean of 0 and standard deviation of 1,000.

After approximating the model parameters' posteriors with the distribution of the Gibbs samples, we investigate their significance. Based on the 95% credible intervals of our parameters, we conclude that all of the intercepts and the vast majority of our time trend in Model 1 are significant. Figure 4 maps the time trend, δ_i and shows the crime rate changes at different rates over time for each zipcode in Chicago. The map indicates that overall crime rates are decreasing in all areas of Chicago since δ_i's are all negative. In fact, the posterior mean of the grand time trend $\delta_0 = -0.066$ signifies that on average, for each additional year, the odds of a Chicago resident being involved in a crime incident decreases by 6.3% ($= 1 - e^{-0.066}$). (Odds is the ratio of probability of being involved in a crime incident over one minus that probability.) The smallest decrease in odds of crime is 3.5% and largest decrease is 8.5%. The crime hot spots, areas with small decrease in odds of crimes, are highlighted in bright orange. In these hot spots, crime is not decreasing over time as much as in other areas, particularly areas colored in dark blue and purple.

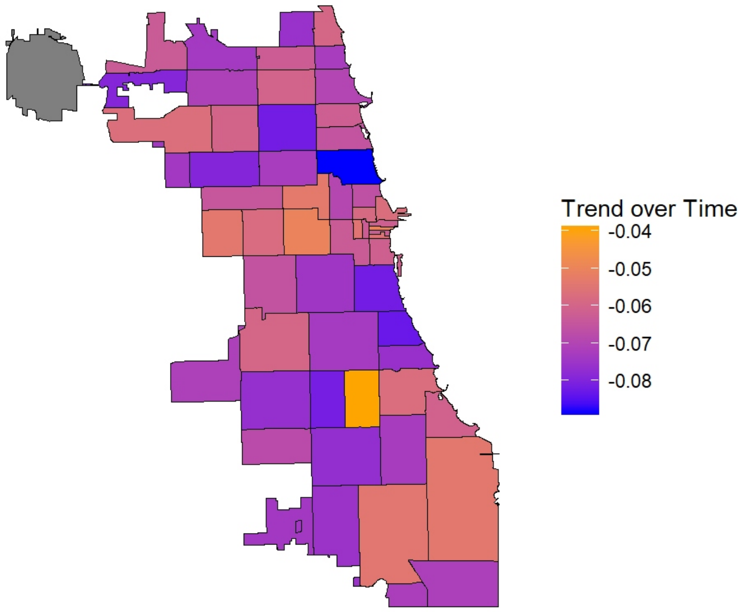

Figure 4: Differential crime growth by zipcodes

Since Chicago is particularly well-known for its high level of violent crime including homicides, we divide our data into two groups-violent and non-violent crimes-to examine whether hot spots for these two groups vary. We classify the following types of crime as 'violent': robbery, battery, burglary, assault, homicide, sex offense, criminal sexual assault, and arson. All other crime types in our data are considered 'non-violent'.

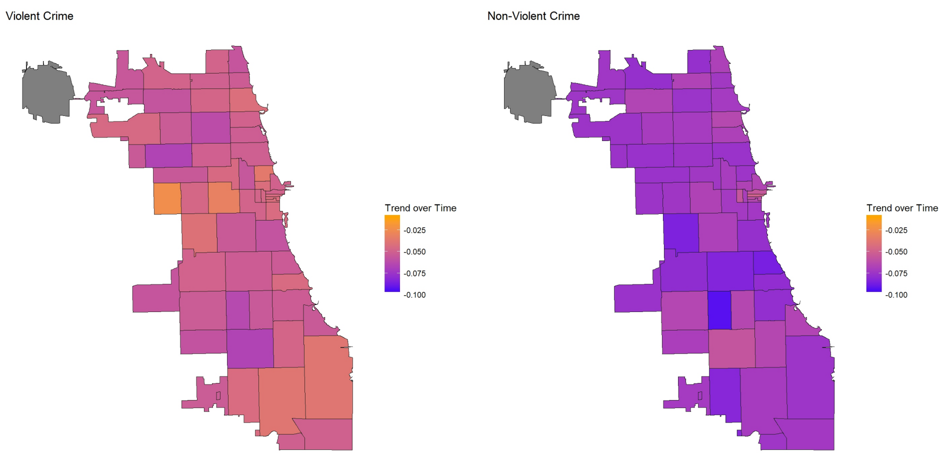

Figure 5: Differential crime growth by zipcodes and types of crime

Figure 5 above indicates that both violent and non-violent crimes in Chicago are decreasing over time. However, non-violent crimes in Chicago decrease over time by a greater extent than violent crimes. This is indicated by a darker shade of purple for non-violent crimes in Figure 5. Specifically, there's a 95% chance that odds of non-violent crime decreases by 6.9% ($\delta_0 = -0.071$) each year compared to a decrease of 4.8% ($\delta_0 = -0.049$) for violent crime in Chicago in general.

3.2 Incorporating Demographics

In Figure 6, we plotted the log of per capita income against the log of crime rate by percentage of white population in each zipcode. In areas with large white populations, an increase in the per capita income seems to result in greater crime rates. However, when an area has a greater non-white population, the crime rate seems to fall as per capita income increases. Thus, we hypothesized that the relationship between income and crime depends on an area's demographic population. For this reason, we derived Model 2 by taking into consideration income and the percentage of white people in each area.

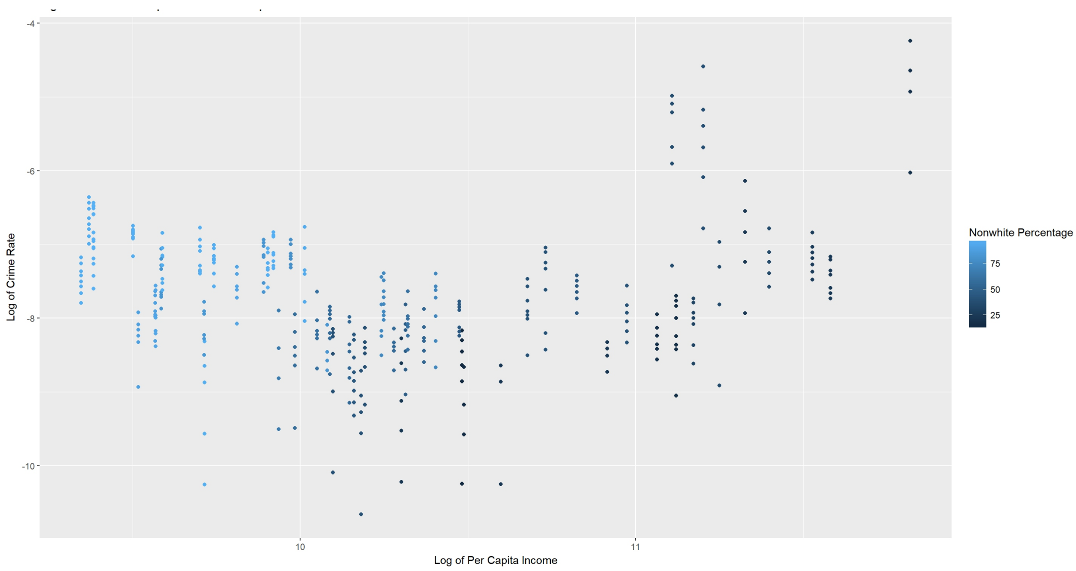

Figure 6: Relationship between income, percent white population, and crime rate

In our second model, we examined changes in the crime rate while taking into account differences in race and income levels in specific areas. Model 2 resembles Model 1, except it takes into consideration income and demographic information for each zipcode. As shown in Figure 6, the relationship between income and crime rate seems to be different for areas with varying levels of white populations. For instance, while crime rate seems to decrease as income increases in zipcodes whose percent white population is below 20%, it seems to increase in those with above 75% white population. Therefore, we decided to discretize our demographic variable into five groups depending on the percentage of white people living in each zipcode: more than 80%, between 60-80%, between 40-60%, between 20-40%, and less than 20% white percentage. Each group contains 4, 12, 14, 7, and 21 zipcodes respectively.

$$y_{ij} \mid p_{ij}, \beta_0, \beta_i, \delta_0, \delta_i, \sigma_e, \tau_0, \tau_1, \tau_2 \sim Bin(p_{ij}, n_{ij})$$

$$log(\frac{p_{ij}}{1-p_{ij}}) \mid \beta_0, \beta_i, \delta_0, \delta_i, \sigma_e, \tau_0, \tau_1, \tau_2 = \beta_i + \delta_i * time_j$$

$$\beta_i \mid \beta_0, \sigma_e, \tau_0, \tau_2 \sim N(\beta_0 + \sum_{e=1}^{5} \sigma_e * income_i * perwhite_e, \tau_0)$$

$$\delta_i \mid \delta_0, \tau_1 \sim N(\delta_0, \tau_1)$$

$$\beta_0 \sim N(0, 1000^2)$$

$$\delta_0 \sim N(0, 1000^2)$$

$$\sigma_e \mid \tau_2 \sim N(0, \tau_2)$$

$$\tau_0, \tau_1, \tau_2 \sim Gamma(0.5, 0.0005)$$

where $i = \{1, ..., 58\}$, $j = \{1, ..., 5\}$, $e = \{1, ..., 5\}$

We "dummy"" coded the discretized percent white population variable described above as perwhite$_e$ where $e \in 1, 2, 3, 4, 5$. The parameter, σ_e, allows income to have a relationship with the log odds of crime for each zipcode group. We decided to include this term in the prior for β_i because each zipcode is nested in one of the five groups. Thus, each zipcode's base log odds of crime is normally distributed around the grand intercept corrected for the group's association with income. We also found that our Markov Chains converged better with this parametrization than when σ_e were included in the log odds of crime level. In addition, we provided vague priors for all the parameters, similar to Model 1, due to limited prior knowledge. Based on the 95% credible intervals of our parameters, all of the intercepts and the vast majority of the time trend parameters in Model 2 are significant. Only σ_2 and σ_5, which describes log odds of crime trend as income increases for areas with 60-80% and under 20% white population, are statistically significant at the 95% credible intervals.

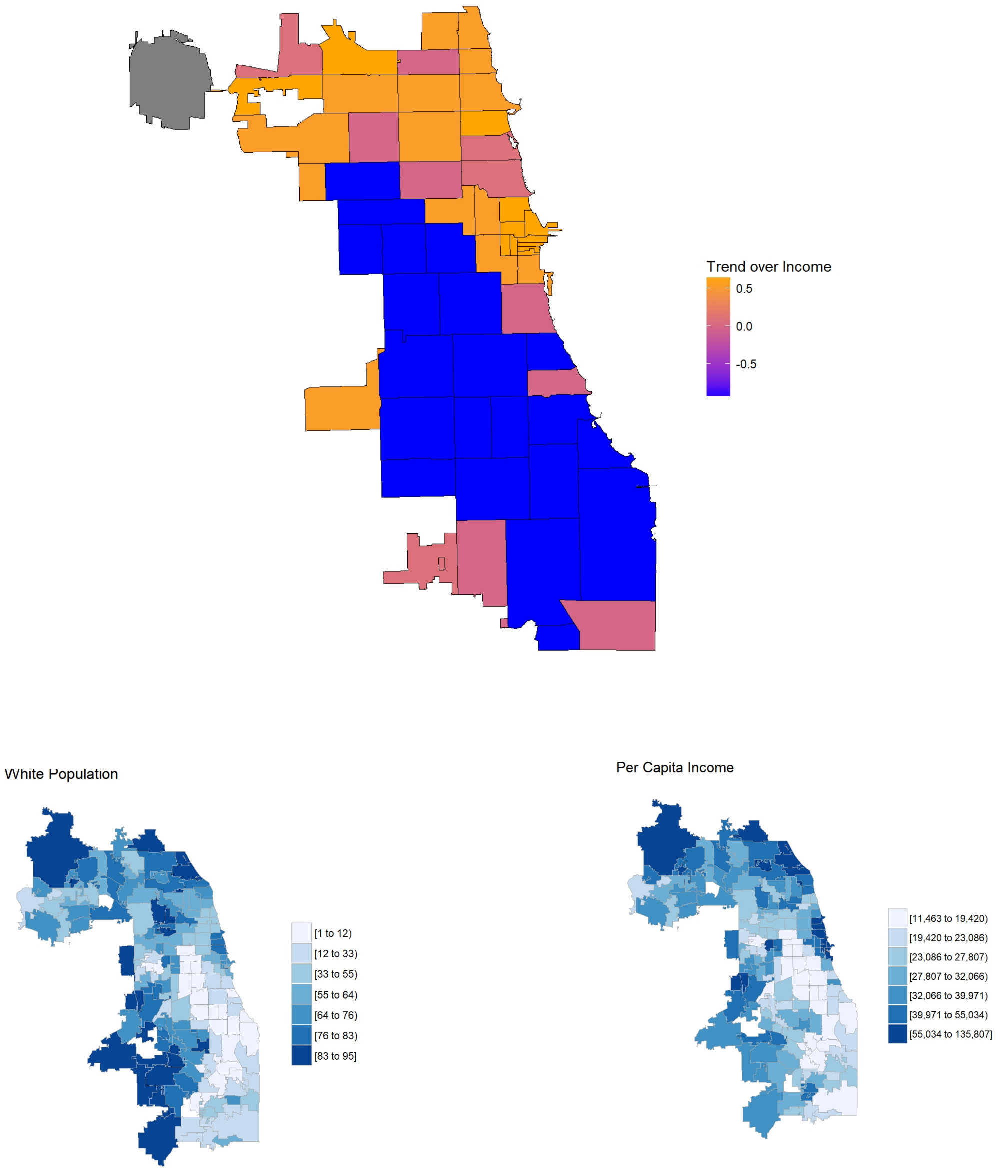

Figure 7: Differential crime growth based on income, percent white population and zipcodes (top). We can see correlations between crime trend and distribution of white population (lower left) and per capita income (lower right).

Figure 7 maps σ_e, which represents the relationship between crime rate and income. The area highlighted in blue represents areas with less than 20% white populations. In these areas, the number of crime decreases over time as per capita income increases ($\sigma_5 = -0.939$). The bright orange areas represent areas with between 60-80% white populations. In these areas, crime increases over time as per capita income increases ($\sigma_2 = 0.633$). This map follows closely to Figure 6, which indicates a negative relationship between crime rate and per capita income in areas with large non-white populations.

Figure 8 examines the crime differential growth based on income and demographics for both violent

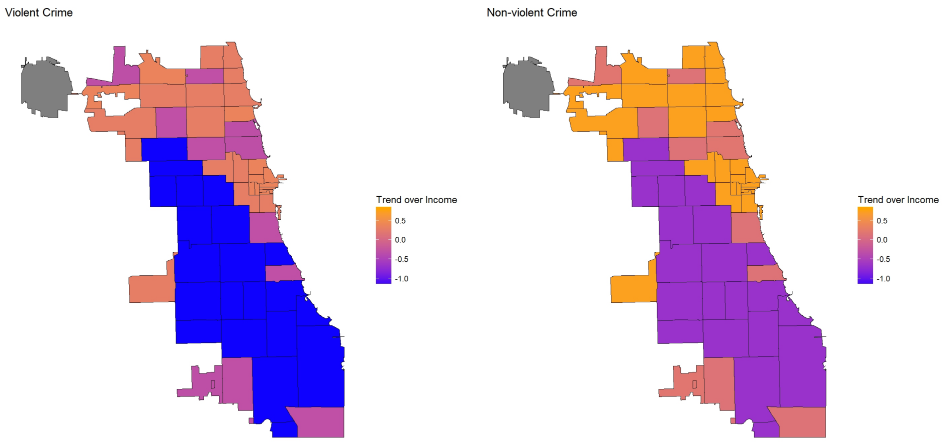

Figure 8: Differential crime growth based on income, percent white population, and zipcodes for violent (left) and non-violent (right) crimes

and non-violent crimes. According to the maps, the number of violent crime increases over time as per capita income increases for areas with between 60-80% white populations, as indicated in orange ($\sigma_2 = 0.365$). In areas where the white population falls below 20%, the number of violent crimes decreases as per capita income increases, as highlighted in blue ($\sigma_5 = -1.192$). This indicates that areas with large non-white populations face lower violent crimes as income increases whereas areas with larger white populations face the opposite effect with more income. The trend for non-violent crimes is similar to the trend in violent crimes. However, in areas with less than 20% white populations, crime rates are not decreasing as much as violent crimes when income increases (for non-violent crime, $\sigma_5 = -0.6875$). In addition, in areas with between 60-80% white populations, non-violent crime is increasing to a greater extent than violent crime as income increases (for non-violent crime, $\sigma_2 = 0.738$). As mentioned above, these two regions are the only groups with statistically significant parameters.

As an alternative to our Model 2, our last model, Model 3, also takes into account the interaction between income and race. However, instead of discretizing the demographic variable (i.e. percent white), we kept the variable as continuous and modeled crime rate with an interaction term between income and the percentage of white populations.

Since the correlation between income and the percentage of white population is high, with a correlation of 0.72, we corrected for the collinearity between the two predictors. We orthogonalized the predictors by creating a new predictor, $perwhite^+$. This predictor represent the portion of the original predictor that cannot be explained by the $income$ variable. Similarly, the $(perwhite * income)^+$ variable represents the portion of the interaction term that is not already explained by the $income$ and $perwhite$ variables. We did this by using QR decomposition so the new predictors satisfies the following equations:

$$perwhite^+ = perwhite - w_1 * income$$

$$(perwhite * income)^+ = perwhite * (income) - w_2 * perwhite - w_3 * income$$

$$perwhite^+.income = 0$$

$$(perwhite * income)^+.income = 0; (perwhite * income)^+.perwhite = 0$$

We could get the effects of the original predictors by premultiplying the effects of new predictors by the matrix R^T from the decomposition above. Orthogonalization of correlated predictors has been shown to improve Markov Chains convergence, which we find to be true for our model (Browne et. al. 2009). We also chose to include the parameters $\omega_1, \omega_2, \omega_3$ for effects of income, percent white population, and their interaction term respectively in the prior for β_i. Because β_i represents zipcode level random effect, and $\omega_1, \omega_2, \omega_3$ are city-level average effects, we chose to model β_i as normally distributed around its expected value (given its income and percent white population) estimated with city-level effects. In addition, we saw better mixing and convergence of Markov Chains with this parametrization compared to when we included these parameters in the log odds of crime level. All priors for our parameters are vague, similar to both Models 1 and 2.

$$y_{ij} \mid p_{ij}, \beta_0, \beta_i, \delta_0, \delta_i, \omega_1, \omega_2, \omega_3, \tau_0, \tau_1 \sim Bin(p_{ij}, n_{ij})$$
$$log(\frac{p_{ij}}{1-p_{ij}}) \mid \beta_0, \beta_i, \delta_0, \delta_i, \omega_1, \omega_2, \omega_3, \tau_0, \tau_1 = \beta_i + \delta_i * time_j$$
$$\beta_i \mid \beta_0, \omega_1, \omega_2, \omega_3, \tau_0 \sim N(\beta_0 + \omega_1 * income_i + \omega_2 * perwhite_i^+ + \omega_3 * (perwhite * income)_i^+, \tau_0^{-1})$$
$$\delta_i \mid \delta_i, \tau_1 \sim N(\delta_0, \tau_1^{-1})$$
$$\omega_1, \omega_2, \omega_3 \sim N(0, 1000^2)$$
$$\beta_0, \delta_0 \sim N(0, 1000^2)$$
$$\tau_0, \tau_1 \sim Gamma(0.5, 0.0005)$$

where $i = \{1, ..., 58\}, j = \{1, ..., 5\}$

Based on the 95% credible intervals of our parameters, we concluded that all of the intercepts and the majority of time trend parameters in Model 3 are significant. The mean estimate of ω_1 is -1.94 and the standard error is 0.003. Since ω_1 is considered significant, we conclude that income and crime have a significant and negative relationship. The mean estimate of ω_2 is 7.758 and the standard error is 0.013, meaning crime and the percentage of white population in an area are positively correlated. ω_3 is not considered significant based on its 95% credible interval. This suggests that different percentages of white people in an area have no effect on the relationship between crime rate and income according to this model. In other words, the interaction term is not significant in this model.

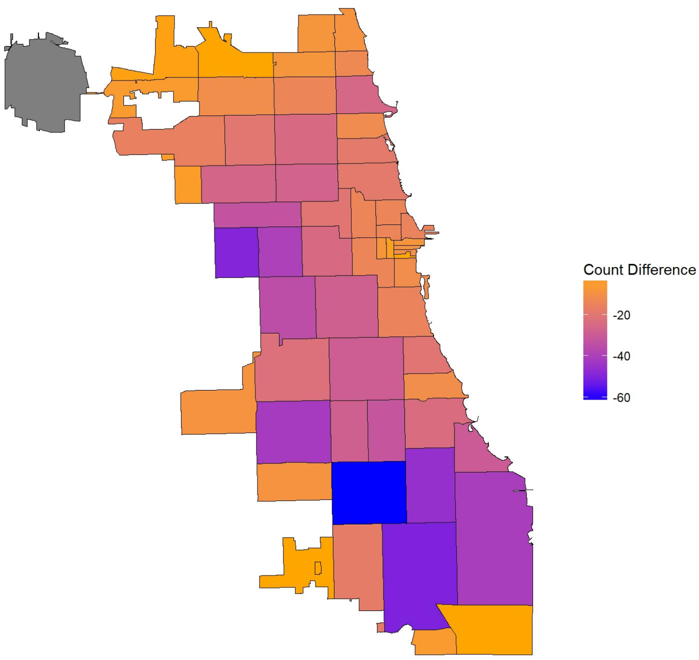

Figure 9: Difference in predicted minus actual crime counts in Chicago in 2016

Finally, we used our last model to predict crime in 2016. Figure 9 maps the difference in crime count between the predicted crime count in 2016 and the actual crime count in 2016. In areas highlighted in dark blue and purple, the predicted crime count is much lower than the actual crime count in 2016. In fact, the difference in crime count is negative in all zipcodes. This suggests that crime in 2016 might be greater than the current trend and therefore, our predictions are underestimating the true crime count.

4 Conclusion

In general, the crime rate in Chicago has decreased in the periods between 2010 and 2015. In addition, crime rates for both violent and non-violent crimes decreased over time but at varying magnitudes. For areas with lower white populations, crime decreased as income rose. For areas with larger non-white populations, crime rate increased as income increased, indicating a discrepancy in crime rates for different demographic areas. Lastly, our third model suggests that crime in 2016 might be increasing to a greater extent than the current crime trend. To conclude, our hot spots analyses and results might be helpful for law enforcement to better understand crime trends in Chicago and predict future crime rates. As a result, it might help with more efficient crime management in specific areas.

5 Limitations and Future Work

One limitation of our research is that we fitted our model using a subset of 10,000 crime observations in Chicago. We would like to validate our result on a larger subset of the original data. In the future, we are interested in examining hot spots for specific types of crime in Chicago. For instance, how do hot spots for homicides compare to hot spots for petty theft? To add, we used zipcodes in our analysis due to its familiarity with general audiences. We could ideally repeat the same analysis using tract data since tracts tend to have more consistent boundaries, similar populations, and is the official unit by which the U.S. Census data are collected. In addition, although we have random effects for each zipcode in our analysis, we can also expect to have autocorrelated random effects between zipcodes that are adjacent. This is because adjacent locations tend to have similar crime rates and behaviors. Therefore, adding a random effect term to account for this correlation can help smooth out dramatic differences in crime rates in adjacent locations. Last but not least, we would like to examine crime rates in specific areas of Chicago in more detail. Certain neighborhoods, such as Fuller Park, experience significantly greater homicide rates compared to other areas (Lucido, 2016). It would be interesting to examine what factors contribute to this drastic difference in crime rates such as race, income, graduation rates, and poverty levels. We could also explore specific areas within the neighborhood, such as parks and schools, to find hot spots in more local areas.

6 Bibliography

1. Braga, Anthony A., Andrew V. Papachristos, and David M. Hureau. “The Effects of Hot Spots Policing on Crime: An Updated Systematic Review and Meta-Analysis.” Justice Quarterly iFirst:1–31. 2012.

2. Browne, William. "An illustration of the use of reparameterization methods for improving MCMC efficiency in crossed random effect models". 2004.

3. Browne, William J., et al. "The use of simple reparameterizaton to improve the efficiency of Markov chain Monte Carlo estimation for multilevel models with applications to discrete time survival models". Journal of the Royal Statistical Society. 2009.

4. "Crimes - 2001 to Present." City of Chicago Data Portal. Retrieved December 17, 2017.

5. Ford, Matt. “What’s Causing Chicago’s Homicide Spike?” The Atlantic, Atlantic Media Company, 24 Jan. 2017.

6. Law, Jane, et al. “Analyzing Hotspots of Crime Using a Bayesian Spatiotemporal Modeling Approach: A Case Study of Violent Crime in the Greater Toronto Area.” Geographical Analysis, vol. 47, no. 1, Oct. 2014, pp. 1–19.

7. Lucido, Gary. “Chicago’s Safest And Most Dangerous Neighborhoods: Homicide Rates.” ChicagoNow, 28 July 2016.