Crime in Chicago 2001-2017

Ryan Rouleau
University of Colorado
Department of Computer Science
Boulder, CO

Kylee Budai
University of Colorado
Department of Applied Mathematics
Boulder, CO

Joe Rickard
University of Colorado
Department of Computer Science
Boulder, CO

PROBLEM STATEMENT AND MOTIVATION

Chicago is a hot spot for unlawful acts. By identifying crime trends in the city, we aim to aid Chicago PD in their efforts to bring Chicago to its full potential. The main question we will answer is ”How do severity of crimes change by location and month/year in Chicago?”. We plan on answering this question by ranking crimes based on severity, bin those severities into three bins, and develop a model fit across Chicago for each month from the year 2001 until the present. Once we have good spatial fits, we will be able to temporally analyze our data. In addition, we will also predict the number of crimes that we expect to fall within each bin. Hopefully, we will be able to find something interesting in this prediction that will help us with the spatial prediction.

PREVIOUS WORK

Crime data in Chicago has been tracked by the Chicago Police Department since 2001. From this data, the FBI and other governmental agencies have identified where and when crime takes place to better allocate their resources. In addition to government studies, third-parties have also used this data to discover other trends.

Last year, the government conducted a study focusing on which types of demographics they could expect to either unlawfully discharge a weapon into someone, or be shot themselves. They figured that although there are a large number of shootings in Chicago, the majority of them are caused by a small group of people. The idea behind this study was to predict who is most likely going to shoot someone or be shot themselves, which would help the authorities know which areas/types of people need more protection. They were able to make predictions that were about 70% accurate using a model that incorporated where the person lived, if they had been to jail before, and if they had been shot before. The data used to conduct this study included 10 more attributes than the publicly available data and can be read about further at this link.

In 2015, a study was done that considered the effects of weather on crime. It was interested in predicting whether crime would be higher or lower than a yearly average given different weather statistics. This study was done by a student at Northwestern who was curious if the weather had a significant enough effect on people’s moods to affect their criminal behaviors. He used a weather database which included parameters such as temperature, humidity, wind speed, etc. In the end, he discovered that the weather attribute which contributed most significantly to crime was temperature which is not entirely surprising. More can be read about this study at this link

3 years ago a study was completed to determine if Chicago’s sports team’s games affected crime in the city. The study found that generally for games involving the city’s teams, crime during the matches dropped by roughly 15%. This drop increased to 25% for more important contests such as the Super Bowl. The study deduced these results by comparing the amount of crime during NFL, NBA, and MLB games to corresponding times with the same day of week, and same month as the original game where there were no matches being played. The study also binned the crimes into general areas to identify if certain crimes were more or less likely to occur during these games. The study found that there was generally an identical drop in all types. The two researchers obtained their data from sports-reference.com (for the dates times of the games), and the identical data set we are using for the crime data. link

Most of the past work we’ve seen looks at the type of crime and correlates it to location. What makes our approach unique is that it looks at the severity of crime (e.g. murder being worse than gambling) by area and tries to predict the change of severity in the future. This will allow law enforcement to plan for the future and better address community issues before they occur.

DATA SET

The data set we used can be found at this link and contains ∼6.2 million rows, and 22 attributes. It includes all of the crime data in Chicago from 2001 to the present. It is updated each week so we will be able to test our temporal predictions against real data by the end of the project. This data includes fields such as when the crime happened, what type of crime it was, the location of the crime, a description of the location, the IUCR code of the crime (which maps to a more specific type of crime), and whether there was an arrest or not.

In order to do severity mapping, we will use the IUCR code to identify crimes. This data set explains each of the codes. It helped us to narrow down severity into more categories than using the primary type would have. For a clear distinction between primary type and IUCR code, an example of primary type is homicide whereas the IUCR codes break that primary type of homicide into first, second, third degree homicide and manslaughter.

For this project, we will consider 6 of the data’s attributes: Location (Latitude/Longitude), IUCR (which specifies type of crime), Date(Month and Year), Location Description, and Arrest. An example of a few lines of our condensed database is given below. We've cut out location description for the table below to fit into the two-column format. Location descriptions describe where the crime took place. For example, STREET, ALLEY, RESIDENCE, etc.

Month	Year	IUCR	Lat	Long	Arrest
08	2008	1330	41.896	-87.630	1
08	2008	1320	41.699	-87.618	0
08	2008	0486	41.763	-87.615	0

TOOLS

The tools that we will use are

- Python
- R
- git
- bash
- AWK

We planned to use a MySQL database to hold and manipulate our data, but we found working with plain CSV's and writing scripts worked better for us. We also use git for version control, Python for the majority of the scripting, R for data visualization, and bash for running our aforementioned Python scripts automatically.

MILESTONES

1. *Early March* - Finish preprocessing.
2. *Before Spring Break* - Generate heat plots for each month with data.
3. *Spring Break* - Construct regressions for crime counts in Chicago.
4. *Mid April* - Build the 3-D model with the month plots and determine a good temporal prediction method.
5. *End of April* - Finish analysis and consider running same analysis for each year.

WHAT HAS BEEN ACHEIVED SO FAR

We've written scripts to preprocess our data, developed heat maps correlating severities of crimes to location in Chicago, and have created regressions to model the counts of crime and predict for the future. A detailed analysis of our work so far is under the "Results so far Section" section.

PREPROCESSING AND SEVERITY MAPPING

In order to perform a sentiment or severity analysis on the data, we needed to consider how to be the least subjective with our data. In order to reduce subjectivity, we removed all crimes with the primary type "OTHER OFFENSE". After performing analysis on the data with these crimes removed, we worked them into our severity ranking system.

We used a severity ranking fro crimes developed in Canada in 2015 to begin constructing a sentiment map. It provided us with a loose structure to map crimes to severities and can be found at this link. Using this main structure, we were able to identify a mapping for a large amount of the IUCR codes. With a little more work, we added in the rest of the crimes, fitting them as best as we could, using data about maximum and minimum prison sentences and fines in order to not be too subjective.

After a preliminary plotting of the data, we found that the heat map did not show any spatial significance which was surprising. Because of this, we rethought the idea of using a severity map by adding in a binning. We put petty crimes in bin 1, non-aggressive crimes in bin 2, and aggressive crimes in bin 3. Petty crimes ranged from gambling to criminal abortion, non-aggressive crimes from concealed carry license violation to low level sexual assault, and aggressive crimes from high level assault to homicide. We will delve deeper into the binnings in the analysis section.

In addition, we decided that if there was an arrest, then it would indicate that a crime was more severe than the same crime without an arrest. For this reason, we decided to separate the data into two parts: crimes with arrests, and crimes without arrests. We also figured that truncating latitude and longitude to three decimal places would allow us to plot the severities while sustaining the structure of the data without losing much of the quality.

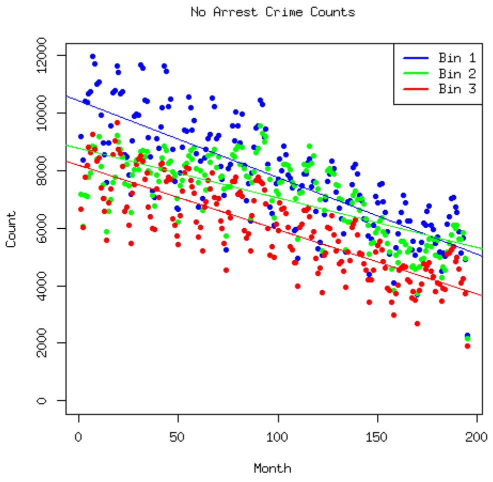
WHAT REMAINS TO BE DONE

We will make our heat maps 3D, improve our regressions for crime counts, and use those regressions to predict crime counts in the future.

RESULTS SO FAR

PREDICTIONS OF CRIME COUNTS

Prior to developing any heat maps, we looked at the number of crimes per bin and ran a simple regression. We did this for each of the three bins and separated them based on whether there was an arrest or not. Below is a plot of the counts of crimes per bin without an arrest. Keep in mind that Bin 1 contains petty crimes, Bin 2 contains non-aggressive crimes, and Bin 3 contains aggressive crimes.



Clearly, the counts on average have been decreasing per month per bin. Minimizing the least squared errors in each of the bins yields a model

$$Y = \alpha + \beta X$$

for each bin where Y is the count and X is the month. The first month, January 2001, was mapped to zero while the last month, December 2016, was mapped to 195. Both coefficients in addition to the standard error of coefficients are given in the table below.

Bin	α	$\hat{\sigma}_\alpha$	β	$\hat{\sigma}_\beta$
1	10418.4	169.6	-26.53	1.50
2	8783.9	140.7	-17.32	1.24
3	8167.9	127.7	-22.22	1.13

With corresponding R^2 values

Bin	R^2
1	0.6182
2	0.501
3	0.6672

which says that the models for bin 1, 2, and 3 account for 61.82%, 50.1%, and 66.72% of the deviation in the data respectively.

In order to verify that these fits are sufficient, we looked at the diagnostic plots for each model and found that, as expected, bin 1 and bin 3 are modeled relatively well with a simple linear regression. The residuals are approximately normally distributed with mean zero, as can be seen with a QQ plot and a plot of residuals versus fitted values. As for the model for bin 2, the diagnostics plot shows many issues in the assumptions of linear regression. The diagnostic plot will be shown below. In order to improve this model, we will attempt to fit the binned data using non-parametric models and run an analysis of variance to compare them.

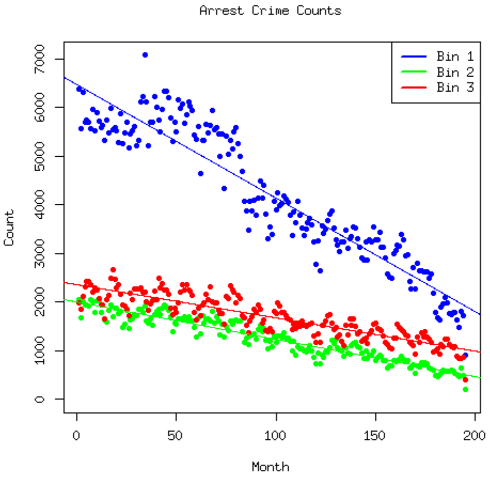
Predictions for the next few months are given in the table below

Bin	January 2017	February 2017	March 2017
1	5218	5191	5165
2	5388	5370	5353
3	3812	3790	3768

with corresponding confidence intervals given by

Bin	January 2017	February 2017	March 2017
1	(4884,5552)	(4855,5529)	(4825,5504)
2	(5110,5665)	(5090.5650)	(5071,5635)
3	(3561,4064)	(3536,4044)	(3512,4023)

Below is a plot of binned crimes where there was an arrest in addition to least squared linear regression fits.



The corresponding coefficients are given by

Bin	α	$\hat{\sigma}_\alpha$	β	$\hat{\sigma}_\beta$
1	6473.6	69.00	-23.3	0.61
2	1995.7	19.46	-7.62	0.17
3	23.52.2	29.7	-6.84	0.26

and R^2 values given by

Bin	R^2
1	0.8833
2	0.9102
3	0.7791

Looking at the diagnostics plots in addition to the R^2 values, these three models explain much more of the variation in the data than the models in the no arrest case.

Predictions for the next couple months are given by

Bin	January 2017	February 2017	March 2017
1	1900	1877	1853
2	502	495	487
3	101	1003	997

with 95% confidence intervals given by

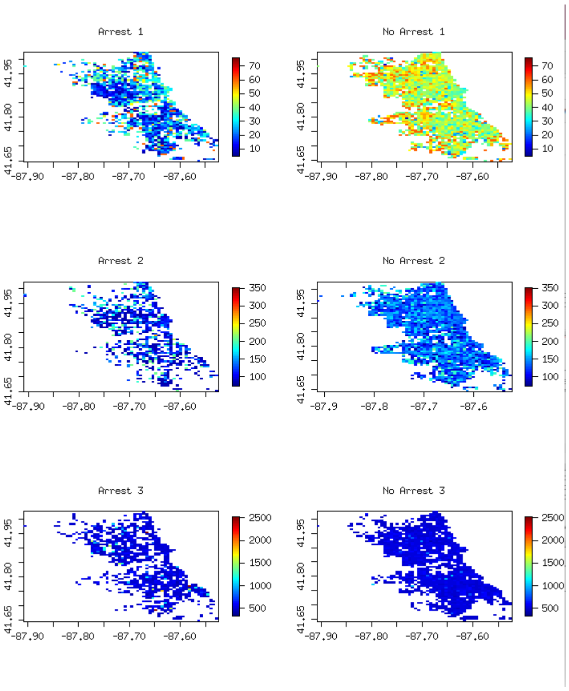
Bin	January 2017	February 2017	March 2017
1	(1764,2036)	(1739,2013)	(1715,1991)
2	(464,540)	(456,533)	(448,526)
3	(951,1068)	(944,1062)	(937,1056)

These models could be improved by considering the months (January, February, etc) as another covariate. This is obvious because of the non-random deviation about the line of fit which appears to be somewhat correlated with time of year more so than was evident from the non-arrest data.

We will run different models on these data and consider time of year as a covariate, hoping to improve prediction.

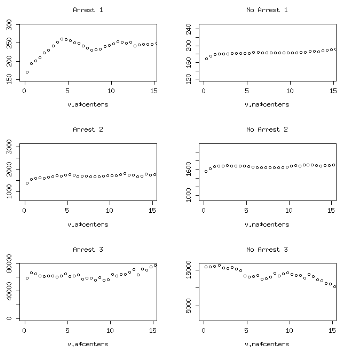
HEAT MAPS

In order to be able to evaluate temporal significance, we considered each month in the data separately. We split the data within each month so that we had test data and training data and could run the following statistical analysis on the training data. We generated six gradient plots per month that show average crime severity per location in Chicago for each of our three bins, arrest and no arrest. We plan to generate a 3D plot over time but for now have two dimensional models that look like this:



The heat maps on the left are for the arrest data and on the right are for the non-arrest data and the three columns are each bins. Notice the different axes, corresponding to the severity levels per bin.

In order to assess spatial dependence, we looked at binned semivariograms which we will explain further in depth later on in the project process. For now, understanding that a semivariogram considers the variance between data points and can detect spatial correlation is sufficient. The following plots are the plots of binned semivariograms for each of the heat maps. These will help us generate complete heat maps by allowing us to determine spatial parameters.



These show that in many of the cases, the majority of the spatial variation is due to noise. This is only for one month (January 2001) and is not a consistent trend across all of the months. However, in the first plot (bin 1 arrest data), there is some clear spatial correlation. This can be seen by the fact that there is a steep incline in the binned semivariogram plot for small distances which will be explained later on.

Once we have a spatial model for each month, we will be able to temporally analyze these models as a unit in a hope that we will be able to make a prediction for the consecutive month. We will run a temporal analysis by layering the monthly spatial predictions on top of each other, forming a three dimensional model. This model will hopefully have an obvious progression of crime which will allow us to develop a reasonable prediction. Once we generate a prediction, we will compare the percentage of estimated high, medium, and low level crimes with those in models that we trust. From these percentages, we will be able to predict how many crimes will fit into each category.

After running the analysis over the months, we may run an analysis over the years if time permits. We would do the exact same thing with the goal of predicting crime in Chicago for 2017.