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 time in Chicago?". We plan on answering this question by ranking crimes based on severity and developing a model fit across Chicago for each month from the year 2001 until the present. Once we have good spatial fits, we should be able to temporally analyze our data.

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. Third-parties have also used this data to discover other trends:

- Predicting demographics of shooting victims (link)
- Predicting crimes in Chicago from weather (link)
- Plotting crimes by location (link)

DATA SET

The data set we will use can be found at this link and contains ∼6.2 million rows, and 22 attributes. We will use 3 of these attributes: date/time, location, and type of crime.

PROPOSED WORK

We will reduce the size of our database by filtering out unnecessary information, leaving only what we care about: date and time, type of crime, and location. We will store these three attributes in a relational database. In order to do any spatial analysis, we will need to encode the type of crime by severity. In order to encode the type of crime in a way that will not skew our results, we will research public sentiments on different types of crimes and rank accordingly. This will be the most important part of the entire project. From here, we will determine whether there are any significant outliers that need to be removed from the data. Since the data is relatively full, we do not need to worry about empty fields.

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.

EVALUATION METHODS

In order to be able to evaluate temporal significance, we will consider each month in the data separately. We will subset the data within each month so that we have test data and training data and will run the following statistical analysis on the training data. We will generate one gradient plot per month that maps crime severity over location. Looking at these plots, we will be able to determine whether the data is spatially correlated. If it is, we will decide whether the spatial process (which is severity of crime) is strictly stationary, isotropic, or neither. Depending on what we find here, we will generate semivariograms to assess the magnitude of spatial dependence, the potential existence of noise, and to determine a relevant covariance function for these data. Once we fit a covariance function, we will be able to use Kriging to predict a relevant spatial model.

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.

TOOLS

The tools that we will use are

- Python
- MySQL
- R
- Amazon Web Services
- git

We plan on using MySQL for our database, git for version control, Python for the majority of the programming, R for data visualization, and AWS to run our code.

MILESTONES

1. *Early March* - Set up database and finish preprocessing.
2. *Before Spring Break* - Generate heat plots for each month with data.
3. *Spring Break* - Determine a reasonable fit for each of the heat plots and fit the entirety of Chicago with those predictions.
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.

SUMMARY OF PEER REVIEW SESSION

From our peer review session, we identified a couple areas of our work we could improve. Firstly, we created a definite question we were going to answer. Secondly, we narrowed down the scope of our project.