# The Scene of the Crime

## Predicting Theft Locations in Chicago

Benjamin Rothschild

**Abstract**

Since January 1, 2016 there have been 128,000 reported crimes, thefts or burglaries in Chicago, a rate of about 500 per day which is 38% of all crimes. Previous research completed in Los Angeles has shown that mathematical models can better predict crime than dedicated crime analysts and by using these models for policing and preventative measures the city was able to decrease theft-related crime by 7.4% in a controlled experiment. In this paper, I perform a similar analysis in Chicago while using novel data sources that represent community input of neighborhood conditions such as 311 calls, datasets that might indicate high-crime areas such as building violations, red light tickets, liquor licenses and how people describe places through tweets. A system like this could be deployed as a real-time tool to help cities deploy their resources for preventative measures and policing.[1]

---

[1] This paper was prepared for MACS 30200 at the University of Chicago. A repository with all the code used to prepare and analyze the data can be found on GitHub here https://github.com/b-nroths/MACS30200proj.

## Introduction

The confluence of the "big data" and "open data" movements has given cities the opportunity to measure and improve the services they provide their citizens. As cities begin to collect and publish data online, researchers are given a new microscope to study and evaluate how cities operate. By combining a city's data with new "big" datasets which provide data at a finer grain and geographic scale than previously available, researchers can analyze cities in new ways. Previous research has shown that combing these datasets can be useful in analysis of city functions such as food and building inspections or property valuation. In the research field of crime prevention, a previous study in Los Angeles showed that a predictive model could outperform a dedicated crime analyst by 40% and by using this predictive model the LAPD was able to reduce theft by 9.8% in a controlled experiment.

In this paper, I contribute to this growing field of research by creating a predictive model to predict the location of crimes in Chicago since January 1, 2016. Crime hotspot maps have long been a tool police departments use to allocate resources. By enriching a crime model with new big data sources, I will investigate how accurate a model of crime prediction can be in Chicago. In my model, I use datasets that represent community input (311 calls), crime magnets (liquor stores, red light tickets), and place descriptions (tweets from locations and neighborhoods). I test my model against different time horizons and different boundary sizes and use an OLS and Spatial OLS model for my prediction.

I find that my prediction model performs best for the Police Beats shapefiles whose average size is .84 square miles. I tested both a OLS and Spatial OLS model and find that they perform with in 5% of one another. The addition of "big" datasets such as Twitter improves the model's performance by almost 5% and demonstrates how "big" datasets can help predict the occurrences of thefts in Chicago.

## Literature Review

Research on crime spans academia, industry, and government. Not only has there been a considerable amount of research done in economics, sociology, and psychology departments but several companies have also provided independent research on the topic and published their

methods online. In addition, city and state governments have used their proprietary and publically available datasets to forecast and optimize their resource allocations to more efficiently distribute their limited resources across their departments. In this literature review I will briefly cover the academic work as well as give an overview of the progress some private companies have made solving similar research questions. Lastly, I will cover some methods that states and cities use to optimized resource allocation in different departments.

Research about how crime rates vary over time and location has been a topic of research for over 150 years. In 1835 Adolphe Quetelet, a Belgian sociologist and mathematician, analyzed crime statistics in France which were first tracked and published in 1827. Using mathematical models, Quetelet found that some people were more likely commit crimes than others such as younger, poorer, and unemployed.[2] In the economics literature, the study of crime began with Becker's model of rationality and criminal activity. Becker's model states that crime comes from "the choices that reasoning individuals make between criminal and non-criminal courses of action, choices informed by predictions of the likely merits and demerits of available alternatives".[3] Further study using this model has been done around deterrence methods such as how capital punishment has decreased homicides in the United States.[4] Another seminal article in the economics literature by Ed Glaeser studied why crime varies so much over time and place. In his paper, he argued that the variation in social interactions create enough variability between individuals to explain the high variance of crime rates.[5] In this model, the amount of social interaction is inversely related to the degree of the crime which means that social interactions are higher in minor crimes and lower in more serious crimes like homicide. This research theory relates to one of the hypotheses I will look at in my paper, namely, the importance of spatial variables with the occurrences of crimes.

---

[2] Quetelet, Adolphe. *Sur l'homme et le development de ses facultés ou essai de physique sociale*. Vol. 1. Bachelier, 1835.

[3] Becker, Gary S. "Crime and punishment: An economic approach." *The Economic Dimensions of Crime*. Palgrave Macmillan UK, 1968. 13-68.

[4] Ehrlich, Isaac, and George D. Brower. "On the issue of causality in the economic model of crime and law enforcement: Some theoretical considerations and experimental evidence." *The American Economic Review*77.2 (1987): 99-106.

[5] Glaeser, Edward L., Bruce Sacerdote, and Jose A. Scheinkman. "Crime and social interactions." *The Quarterly Journal of Economics* 111.2 (1996): 507-548.

The rise of Big Data has produced a new wave of research on the study of crime and urban life. Whereas previously research relied on highly aggregated statistics and surveys, new data sources such as sensors and digitized records have produced more rich data sources that can be used to analyze classic questions in urban social science research in new ways. For example, some researchers were successfully able to predict the median income of residents in New York City by training a computer vision model on Google Street View images.[6] In addition many private companies have rich data of how people live and transact in cities. Websites like Zillow and Craigslist have data on home prices and other economic transactions while review sites like Yelp and Foursquare has been used to predict foot traffic in restaurants and stores. In addition, the Open Data movement has also been invaluable to the study of cities. Many cities and have started to publically release datasets online to be used for analysis by businesses and researchers. By combining these data sources with city records new areas of research is possible.

One example of this phenomena is that researchers used text analysis on Yelp reviews to predict if a restaurant had a health violation. They matched 1,756 restaurant inspections with reviews on Yelp from 2006 to 2013 and used a SVM model with 10-fold cross validation to predict what restaurants were more likely to have food violations. They could successfully predict 82% of severe offenders from establishments without a violation.[7] In Chicago, data scientists in the city government created a prediction model to more efficiently distribute food inspection officials to food establishments across the city. There are over 15,000 food establishments that are subject to sanitation inspections and only three dozen food inspectors. To better allocate these inspectors to restaurants, data scientists collected data on complaints and matched them to the food establishments business characteristics. This model found that food inspection outcomes were correlated with many factors such as sanitation complaints, nearby burglaries, length of time since the last inspection, and if the establishment had previous critical or serious violations. By using this model to allocate their food inspectors, the city improved the time to discovery of critical violations by 7 days.[8] Similar to these approaches, my research aims to help cities better

[6] Glaeser, Edward L., et al. "Big data and big cities: The promises and limitations of improved measures of urban life." *Economic Inquiry* (2016).
[7] Kang, Jun Seok, et al. "Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews." *EMNLP*. 2013.
[8] https://chicago.github.io/food-inspections-evaluation/

allocate their resources by using novel "Big Data" data sources such as using tweets to measure how people describe places in a city.

In the research area of crime detection and predictive policing, the biggest study to date looked at how effective predictive models were at lowering crimes in Los Angeles and Kent (United Kingdom).[9] Their hypothesis was that a predictive model could be more effective at distributing police resources than a dedicated crime analyst and the improved resource allocation would reduce crime. They created a model to predict crime hotspots in Los Angeles focusing on burglary, car theft, and criminal damage which comprise about 55% of crime volume. They used an epidemic-type aftershock sequence (ETAS) model that estimates the risk associated with both long-term hotspots and short-term models of near-repeat risk. From the results of this model they then conducted two controlled experiments with a division of the Los Angeles Police department and Kent Police Department. In the experiments the ETAS algorithm was put head-to-head with hotspot maps procures by a dedicated crime analyst. Compared to a dedicated crime analyst the predictive model could predict crime successfully at a rate of 9.8% and 6.8% compared to 6.8% and 4.0%. The next step in their research was to use this predictive model to control where police go to measure if the improved policing coverage could lower crime rates. They conducted a controlled experiment and found that on average when using the ETAS forecasts there was a 7.4% reduction in crime volume as a function of patrol time, whereas patrols based upon analyst predictions showed no significant effect. The researchers conclude that dynamic police patrol in response to ETAS crime forecasts can disrupt opportunities and lead to real crime reductions.

There is also a small body of research on improving the accuracy of hotspot maps. Since crime hotspot maps are a widely-used method of displaying spatial crime patterns and allocating police resources, researchers have developed a methodology to better predict crime hotspots. They argue that current hotspot maps often utilize too little of available data and fail to capture short-term changes in risk. They improve on these methods by using an expectation maximization algorithm which can easily be deployed on a desktop computer connected to a police agency

[9] Courneya, Kerry S., et al. "*Randomized controlled trial of exercise training in postmenopausal breast cancer survivors: cardiopulmonary and quality of life outcomes." Journal of clinical oncology 21.9 (2003): 1660-1668.*

Risk Management System. The researchers applied this methodology to homicide and gun crime in Chicago and make their results and methodology available online. [10]

In addition to academic research there have been several private companies that offer software solutions to police departments and thus contributed to this research topic. Often these companies will describe their methods online or publish to academic journals. One example is CivicScape, a company founded by University of Chicago professors and alumni. They leverage information from historical crime-trends, on-the-ground intelligence from police officers, and input from the community and provide a deployable solution to police departments. The input datasets they use include recent crime activity, 311 calls (community input), census tract data, weather forecasts. They use an ensemble of feed-forward neural networks tuned to the specific crime type and location in the city and produce predictions on a three-block radius for every hour. They have open sourced their methodology and data sources on Github.[11] Another company PredPol provides similar predictive models to departments in the United States. They use three data points, crime type, crime location, and crime date to provide agencies with custom crime predictions, usually pinpointing areas within a 500 foot by 500 foot area. They update their models every 6 months and are aimed at supporting dedicated crime analysts in making resource allocation and patrol decisions. They have published their results in several academic papers.[12]

Other Police departments have already deployed real-time analytics systems to improve their policing efforts and trust within the communities they serve. For example, in New York City, the Police Department has created a new system to incorporate real-time feedback into their policing efforts. They will send short surveys out around the clock to over 50,000 smartphone applications by buying mobile advertisements geo-targeted to users who live in the City. In the survey, they will ask questions such as "Do you feel safe in your neighborhood?", "Do you trust the police?", and "Are you confident in the New York Police Department?" The police

---

[10] *Mohler, George. "Marked point process hotspot maps for homicide and gun crime prediction in Chicago." International Journal of Forecasting 30.3 (2014): 491-497.*
[11] *https://github.com/CivicScape/CivicScape*
[12] Mohler, George O., et al. "Self-exciting point process modeling of crime." *Journal of the American Statistical Association* 106.493 (2011): 100-108.

department uses these responses to adapt their block-by-block deployments and tailor their strategies to better serve the communities they police.[13] While this system is new, the police department hopes that it will improve crime prevention and police trust by allowing the department to better target strategies on a block-by-block level. Similar to this system, my research paper investigates how a crime could be predicted on granular level which could be deployed as a real-time analytics system.

**Data**

The data I use in my analysis was collected from publically available data sources. The data sources I use are broadly divided between community input datasets (311 calls), crime magnets (events that correlated with crime), and place descriptions (how people describe places). I was careful to choose datasets that did not have any racial or demographic bias and chose not to include any demographic information in my model. Because this is intended to be a deployable analytics system, transparency about the data and methodology are important. The code I used to extract, prepare and analyze is publically available on GitHub at https://github.com/b-nroths/MACS30200proj.

Most of the community input and crime magnet datasets come from Plenar.io. Plenar.io takes data that cities produce and makes it available through a REST API. In addition to providing data from multiple city portals through a single API, Plenar.io unites all its datasets on a single spatial and temporal index, simplifying the data gathering and cleaning process. The project is funded by the National Science Foundation Computer and Information Science and Engineering directorate though a grant to the Urban Center for Computational Data at the Computation Institute of the University of Chicago and Argonne National Laboratory.[14] By using the same temporal and geographic index across datasets Plenar.io makes it easy to combine datasets from different data portals into one analysis. Most of the data they make available is data published by the City of Chicago through their online data portal.[15] In addition, I collected data from

[13] This system was deployed in early 2017 and described in the New York Times https://www.nytimes.com/2017/05/08/nyregion/nypd-compstat-crime-mapping.html

[14] The NSF grant is described here: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1348865&HistoricalAwards=false

[15] https://data.cityofchicago.org/

Twitter that was geo-tagged to a location in the Chicago City limits. For each dataset, an observation has an attached timestamp and latitude and longitude which will be used to place each event into a spatial region for the spatial regression in my analysis. A table summarizing each variable is below.

**Table 1:** Data Descriptions

| | # Obs. | Avg # Obs/day | Source |
|---|---|---|---|
| Crimes - Theft | 215,075 | 425 | Chicago Police Department |
| Crimes - Non-Theft | 128,487 | 253 | |
| Building Violations | 165,681 | 324 | Chicago Department of Buildings |
| 311 Graffiti Request | 164,318 | 321 | City of Chicago 311 Requests |
| 311 Sanitation Request | 28,537 | 55 | |
| 311 Alley Lights Out | 44,166 | 86 | |
| 311 Vacant Building - Gangs | 2,243 | 4 | |
| 311 Vacant Building Out - No Gangs | 4,333 | 8 | |
| Food Inspection - Pass | 16,945 | 33 | Chicago Department of |
| Food Inspection - Pass w/Condition | 5,674 | 11 | Public Helath Food |
| Food Inspection - Fail | 3,342 | 6 | Protection Program |
| Red Light Tickets[1] | 86,817 | 964 | Chicago Tribune |
| Liquor Licenses[2] | 4,541 | n/a | Department of Business Affairs and Consumer Protection, City of Chicago |
| Tweets - Good Sentiment[3] | 1,052 | n/a | Twitter |
| Tweets - Bad Sentiment | 183 | n/a | |

1. Red light ticket data is only available from 12/2/2013 to 1/3/2014
2. Liquor Licenses dataset is for current liscenses only so it is not shown over time.
3. Twitter dataset was added recently so there is no historical data past what was collected.

**Chicago Crime Dataset 2001 – present**

This dataset includes reported incidents of crimes (with the exceptions of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In many cases addresses are shown at the block location and specific locations are not identified. There are over 400 different codes used to specify a crime divided among 33 primary types. To simplify my regression analysis, I divided the crimes into two types Thefts and Non-Thefts. Thefts include burglaries, robberies, and motor vehicle thefts while non-thefts include violent crimes, battery, assault, criminal damages, among others.[16]

**Building Violations**

[16] A full list of crime types and their descriptions can be found here, https://data.cityofchicago.org/Public-Safety/Chicago-Police-Department-Illinois-Uniform-Crime-R/c7ck-438e

Building violations are issued by the Department of Buildings and include violations such as failed maintenance, infestation, and unsafe conditions among others. For each violation, it is reported when the building failed inspection as well as if it is still in violation of a city code and weather the violation comes from a complaint or periodic inspection.

**311 Service Requests**

311 Service Requests represents a form of community input in my model. 311 requests are non-emergency complaints and notices Chicago residents can make to their city government. They serve as a point of entry for residents, business owners, and visitors to easily access information regarding City programs and the city documents all requests for non-emergency City services and releases data through their online data portal. For my analysis, I use several request types that have the potential to impact crime and/or safety.

**Table 2:** 311 Descriptions

| | Description |
|---|---|
| 311 Graffiti Request | Requests to remove graffiti with city blast trucks |
| 311 Sanitation Request | Complaints such as overflowing dumpsters or garbage in the Alley |
| 311 Alley Lights Out - Gangs | One or more street lights out with reported Gangs or Homeless |
| 311 Alley Lights Out - No Gangs | One or more street lights out without reported Gangs or Homeless |
| 311 Vacant Building | Requests to inspect a vacant building |

**Red Light Tickets**

Red light tickets are compiled by the Chicago Tribune and published monthly. This dataset only includes tickets that are given through the Red-Light Camera system where cameras are installed at intersections and given to people automatically and not by Police giving out tickets.

**Food Inspections**

This data is from inspections of restaurants and other food establishments in Chicago. Inspections are performed by staff from the Chicago Department of Public Health's Food Protection Program using a standardized procedure. Each inspection falls into one of three categories described in the table below.

**Table 3:** Food Inspection Descriptions

| | Description |
|---|---|
| Pass | The establishment meets the minimum requirements of municipal codes and does not have any serious or critical violations |
| Pass with Conditions | The establishment has Serious or Critical violations that are corrected during the inspection or the certified Food Service Sanitation Manager is not present as the time of the Inspection |
| Fail | The establishment has Serious violations that cannot be corrected during the inspection. The business must correct the Serious violations promptly and pass a re- inspection to remain open. Note: the business can also have its license suspended until it passes re-inspection. |

**Liquor Licenses**

Current active liquor and public place of amusement business licenses issued by the Department of Business Affairs and Consumer Protection in the City of Chicago. This dataset includes the when the license was valid from.

**Twitter**

The Twitter dataset was used to get a measurement of how people describe certain places in a city. Since there was no Twitter archive data available the dataset only has data from when I started recording data which was May 20, 2017. To collect the data, I listened to tweets that had a geolocation tagged to them that were within the City of Chicago boundaries. Since tweets can be tagged by City, Town, Neighborhood, or specific location I only included tweets that had a specific location or neighborhood boundary. If tweet had a neighborhood boundary I found the centroid of the neighborhood boundary and considered the tweet to be from that location. To find the sentiment of the tweet I used Google's Cloud Natural Language API which returns a sentiment (from -1 to 1) and magnitude (from 0 to 1) of sentiment for a string of text. I multiplied the sentiment score and magnitude and considered tweets to have good sentiment if they were above a threshold of 0.25 and to have bad sentiment if they had a score of less than -0.25. Some examples of tweets are below.

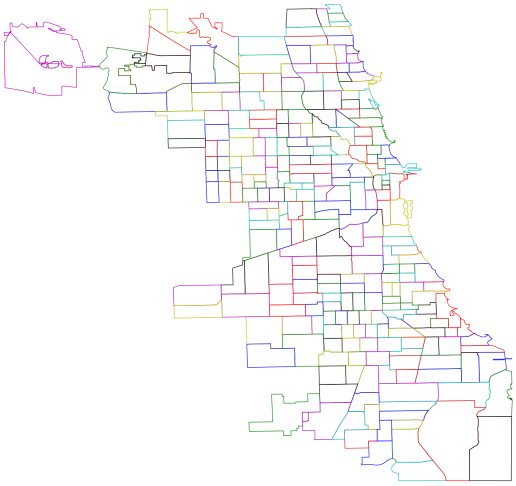
**Table 4:** Tweet Examples

| Score | Category | Tweet |
|---|---|---|
| .55 | Good | I love this place already! - Drinking a Slim Hazy by @mikerphonebrew at @mikerphonebrew https://t.co/8EdB6ERhfJ |
| .72 | Good | Here's our Morton Community who supported our fundraiser event today! THANK YOU all who could not make it but con https://t.co/rKrqMPWf7u |
| -.36 | Bad | I'm really not trying to be in the crib tonight |
| -.35 | Bad | Muffler: $160 Battery: $150 Alternator: $350 Me: Broke as **** |

## Region Boundaries

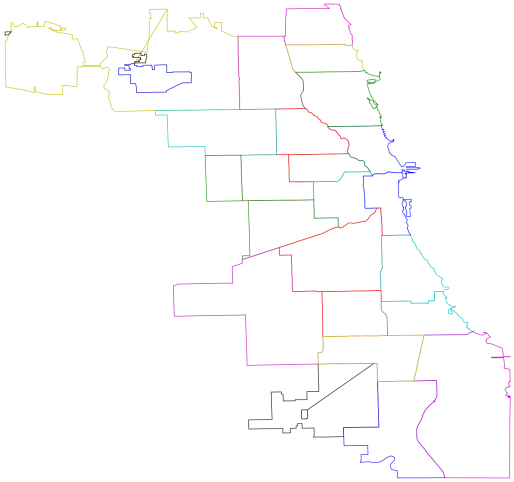
To perform my analysis, I tested my model against multiple region boundaries to see what boundaries gave the best prediction. I identified 8 different possible boundaries which vary in their size and boundary type. A map and description of each boundary tested is below.

Police Beats



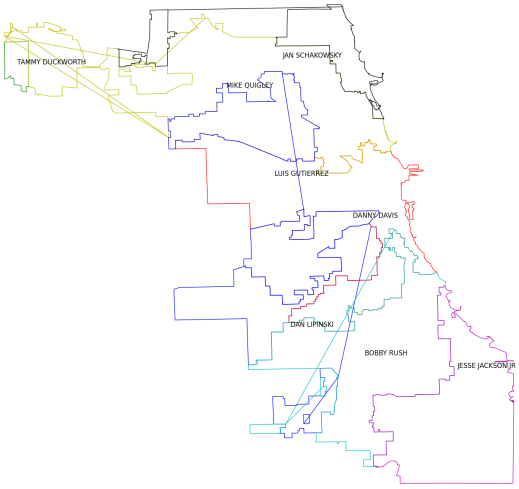
Current police beat boundaries in Chicago.
Source: City of Chicago

Police Districts



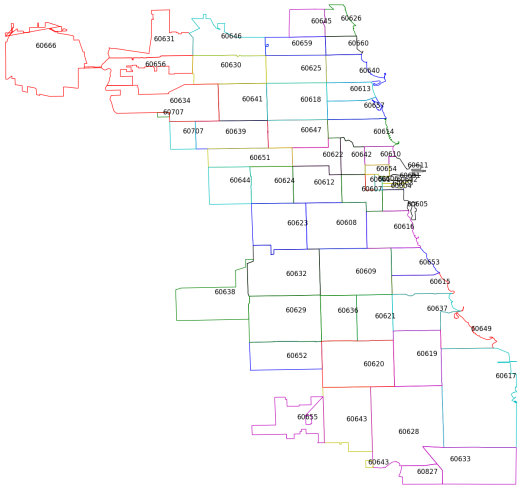
Police district boundaries in Chicago.
Source: City of Chicago

## Congressional Districts




U.S. Congressional district boundaries in Chicago.
Source City of Chicago

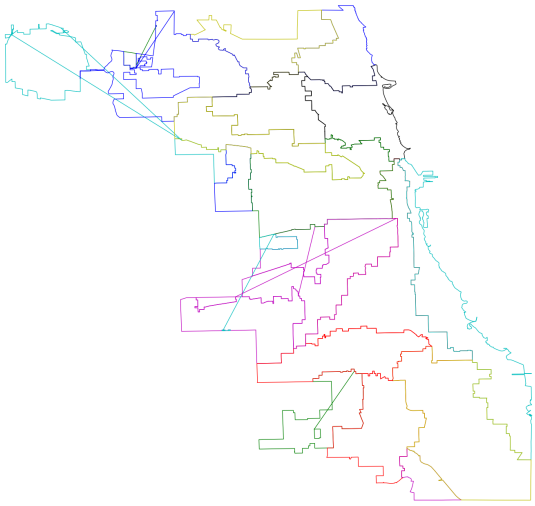
## Zip Codes



ZIP Code boundaries in Chicago.

Source: City of Chicago

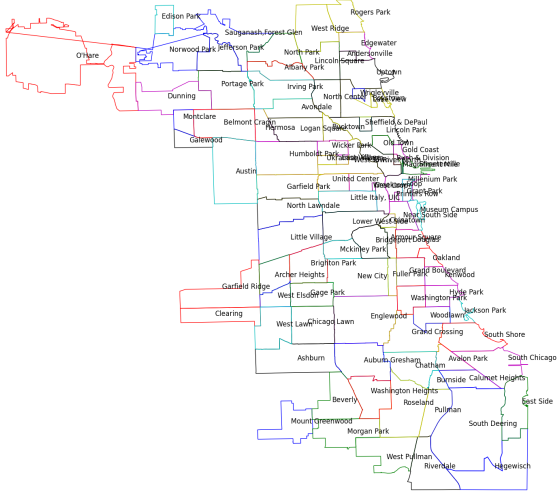
## State Senate Districts



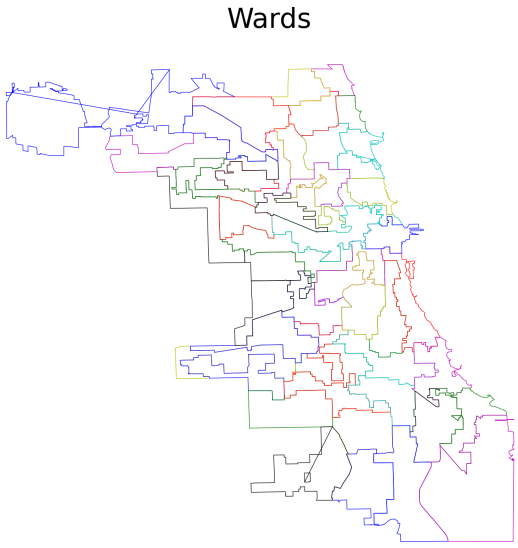
State of Illinois Senate Congressional district boundaries in Chicago.

Source: City of Chicago

## Neighborhoods

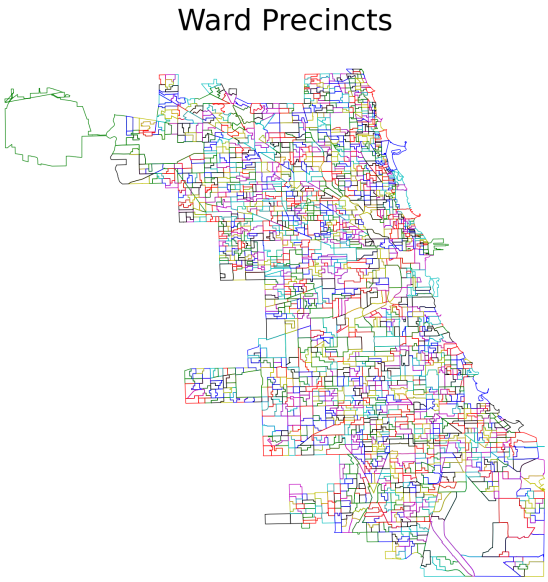


Neighborhood boundaries in Chicago, as developed by the Office of Tourism. These boundaries are approximate and names are not official.
Source: City of Chicago Office of Tourism




Ward boundaries in Chicago.

Source: City of Chicago




Ward precincts, which function as local election districts, in Chicago.
Source: City of Chicago

**Methodology**

The first step I take in my model is to divide all the data points I have into a shape boundary. I do this by loading the data into PostgreSQL with the PostGIS extension and using their point-in-polygon function `st_within()`. This will allow me to get aggregate counts of an event type for a specific shape boundary and time period.

To get a baseline for my prediction I first perform an OLS regression without considering any spatial weights. This baseline will be useful to compare with the spatial regression I perform later as we should expect crime data to have some significant spatial component. Additionally, an OLS regression should give a signal for important variables in the model and their contribution to the crime rate in Chicago. This type of regression is common in analyses of crime rates. The simple OLS regression is below where $X_t$ is an array that includes the count of observations of an event during the period. Some variables have no time component such as # of liquor licenses, number of red lights tickets (data was collected before the time period of my

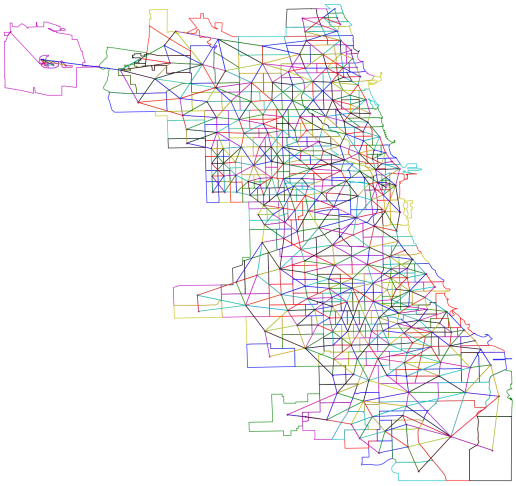
analysis starts), and number of tweets (data was collected after the time period of my analysis ends).

OLS Regression

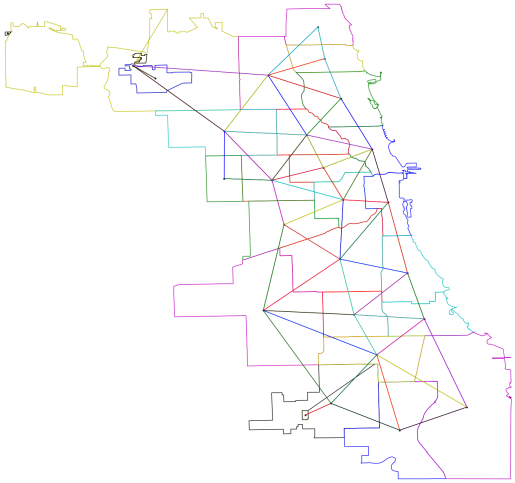
$$Theft_{(t,j)} = \alpha + \beta_1 \boldsymbol{X}_{(t,j)} + \beta_2 \boldsymbol{X}_j + \epsilon_t$$

Next I performed a Spatial OLS regression. The first step in preparing the data is to calculate the queen-contiguity neighbors of each shape. Queen-contiguity is one of several spatial relationships I could have chosen but it is the most common one in spatial regression analysis. Other possible weighting schemes include, rook weights, bishop weights, distances, and k-nearest neighbor weights. Under queen-contiguity, a shape is the spatial neighbor of another if it shares a point or edge with the other shape. To find the neighbors of each shape I used the PySal library Python module with the `queen_from_shapefile()` function.[17] The output is a list of shape ids with their neighbors. Below are the maps of my Chicago boundaries with a line connecting each shape to its queen neighbor.

Police Beats

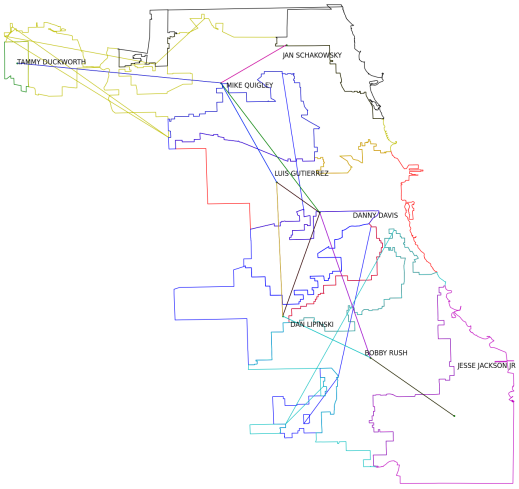


Police Districts

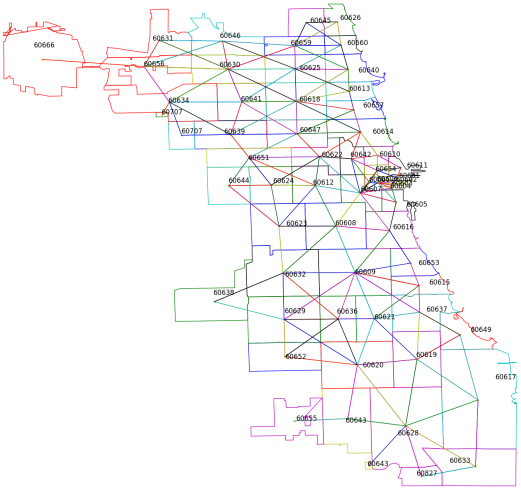


[17] PySal is an open source library (http://pysal.github.io/funding.html) that contains several spatial data anlysis functions.

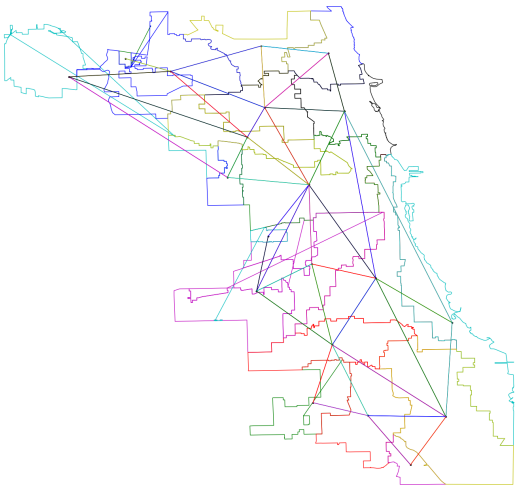
## Congressional Districts



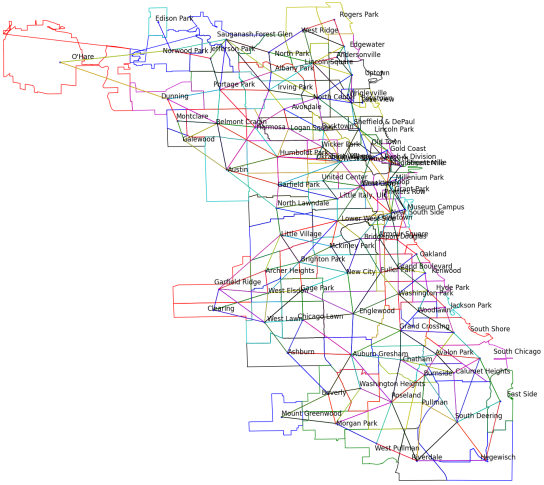

## Zip Codes



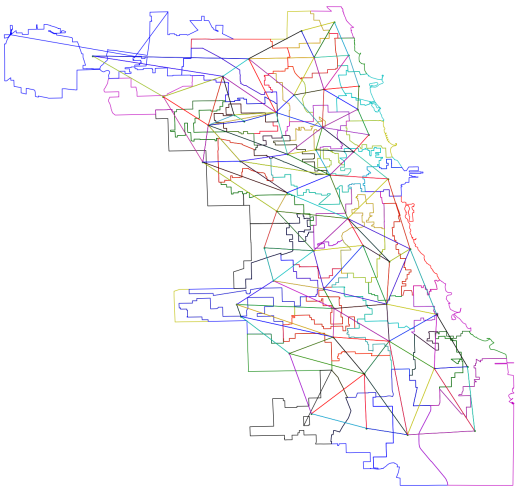

## State Senate Districts



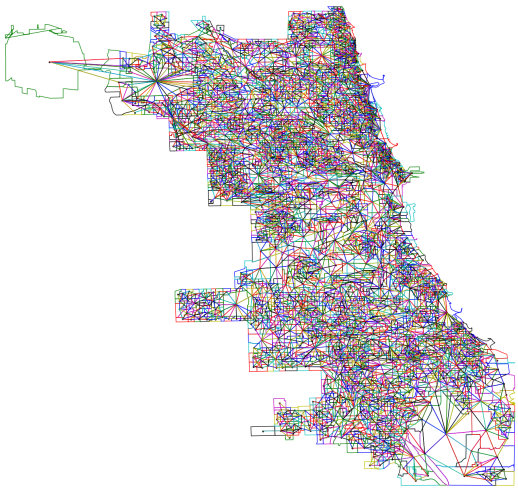
## Neighborhoods



Wards



Ward Precincts



From this new dataset I can now complete a my spatial OLS regression where I average the observations for the neighbors of a shape as describe in the formula below.

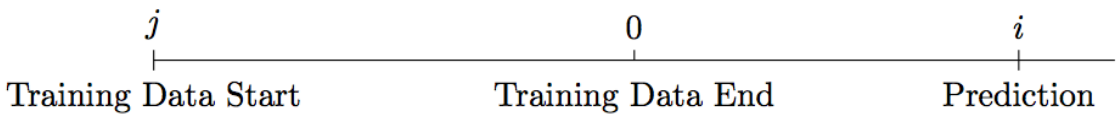
Spatial OLS Regression

$$Theft_{(t,j)} = \alpha + \beta_1 \boldsymbol{X}_{(t,j)} + \beta_2 \boldsymbol{X}_j + \beta_3 \bar{\boldsymbol{X}}_{(t,j_{neighbors})} + \beta_4 \bar{\boldsymbol{X}}_{(j_{neighbors})} + \epsilon_t$$

One last key part of the theory used in my paper is the Cross-Validation measure I use to measure the accuracy of my predictions. Typical cross validation such as k-folds or leave-out-one cross validation would not work in this database because it is important to consider only observations that have happened before the prediction date when training the model. To overcome this I came up with a new cross-validation algorithm I will use:

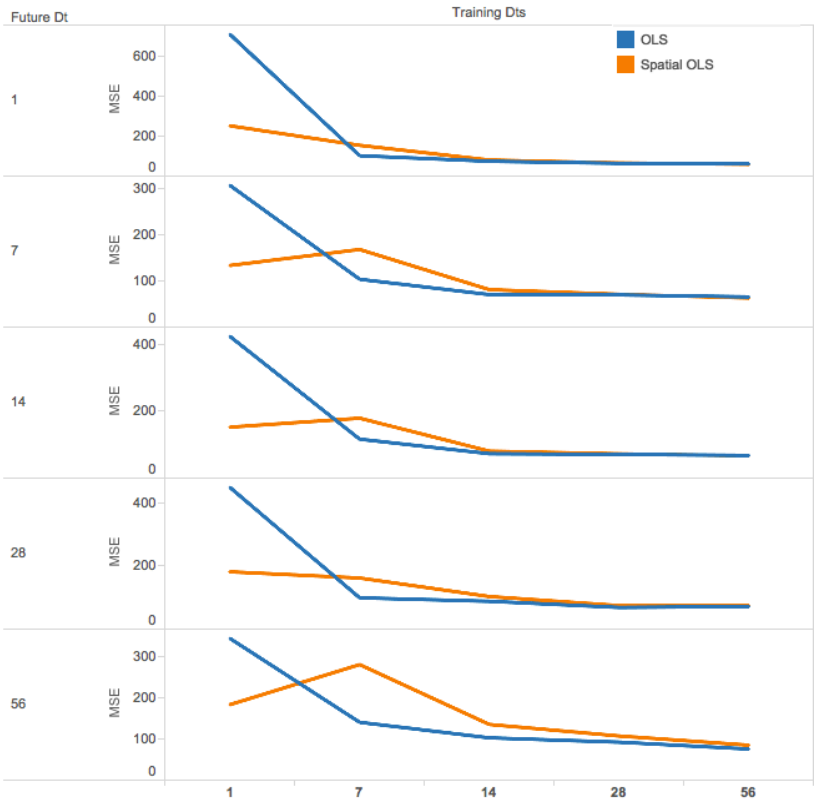
1. Choose a day at random from all possible days with enough data
2. Train the model on the previous ($j$,0)
3. Predict day $i$
4. Average the mean squared error for 200 iterations of this algorithm to report the model's accuracy

A visualization of this timeline is below.

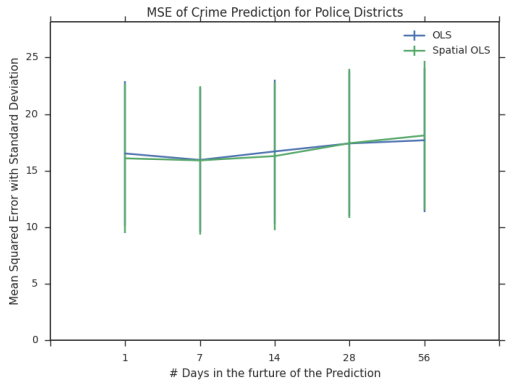


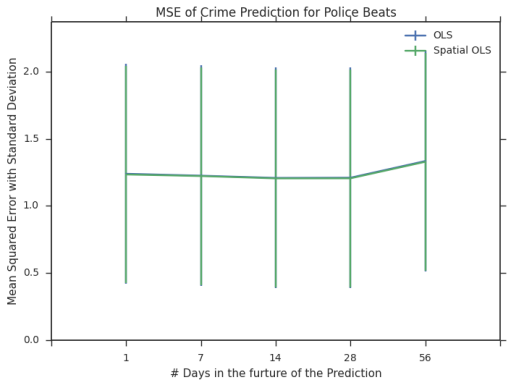

**Results**

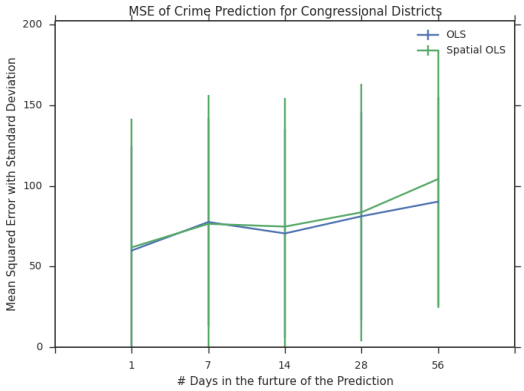
The first value I tested was how much training data I should use in my model. I wasn't sure how much training data would be optimal for a predicting a value. For example, would data from a year ago be as useful in predicting data in the future as data from a few days ago. To test this, I ran my model with different training data start values and varied it from 56, 28, 14, and 7. The results of this analysis is below and show that adding more training data from decreases the average MSE predicted but after about 28 days adding more data does not decrease the model by significantly more. Because of this I decided to make my "knowledge window" 28 days long.

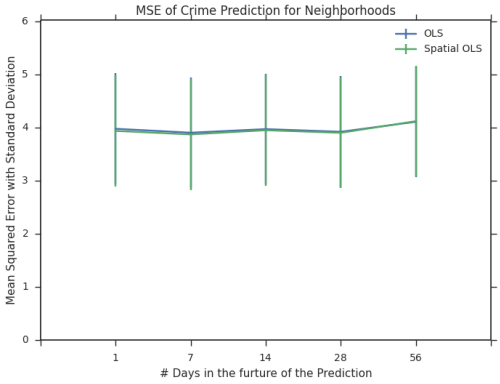


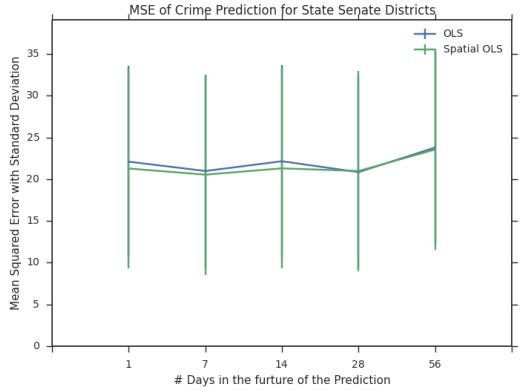
I then performed my Spatial OLS and OLS regression. The results are plotted in the graphs below. I varied how far I was predicting which is shown on the x-axis.

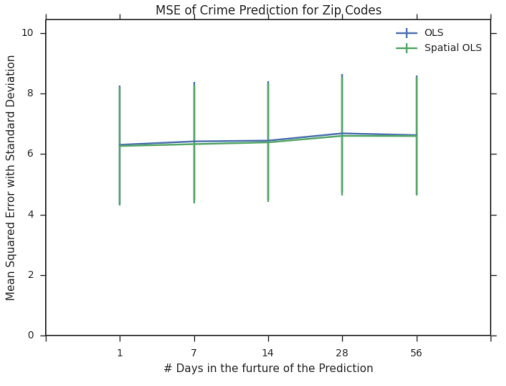


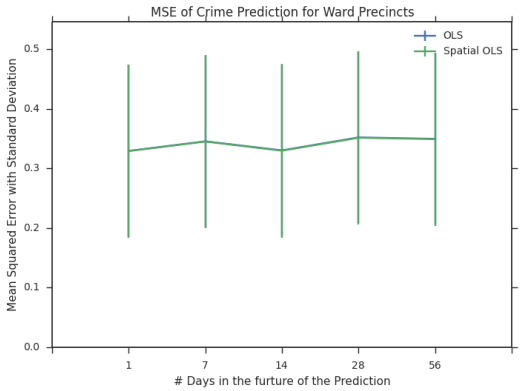



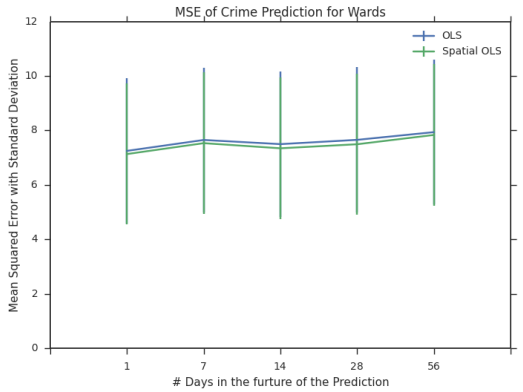



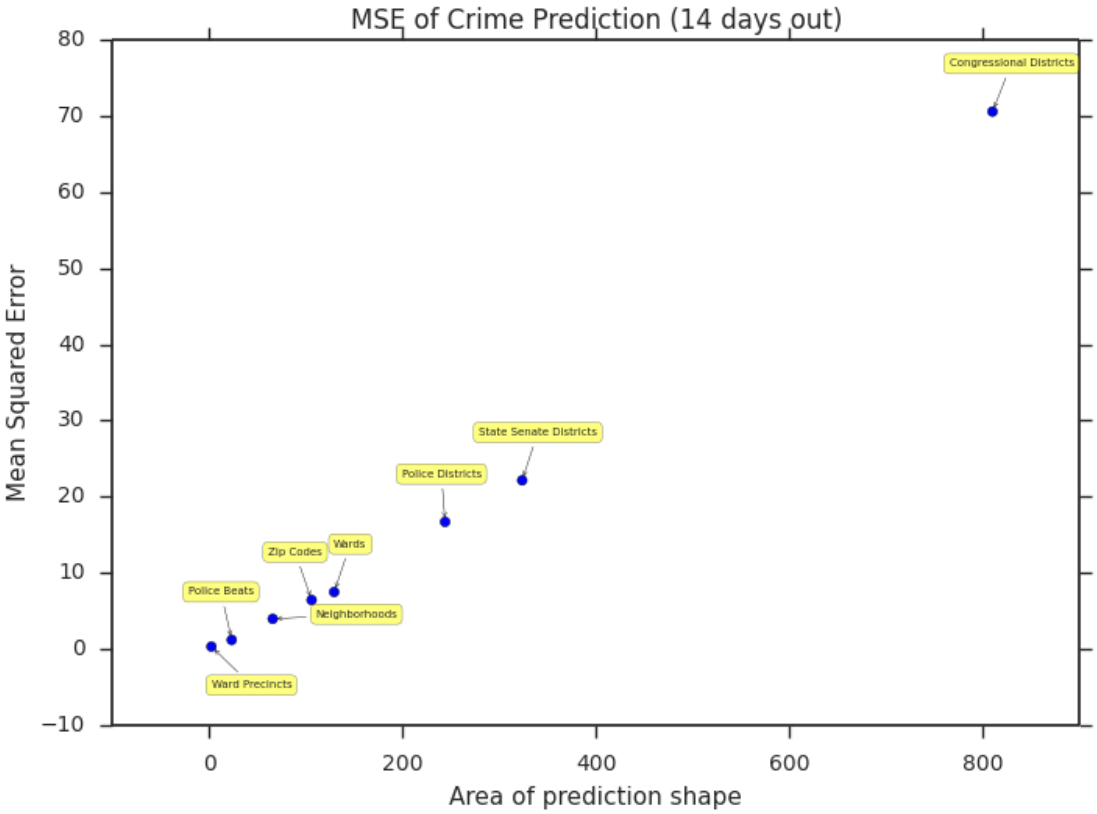





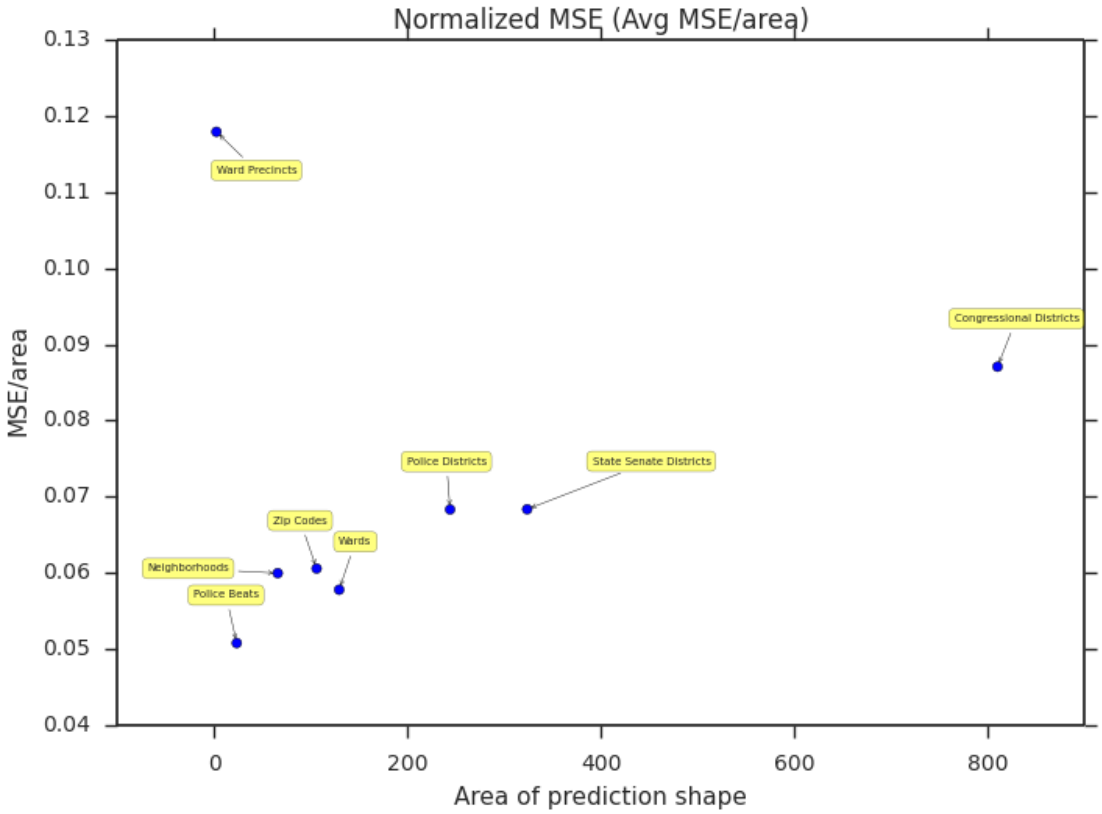

As can be seen in the graphs the difference in prediction accuracy (measured by mean squared error) between the prediction and actual value are not very different between the Spatial OLS and OLS model. However, it is interesting to note which boundaries the Spatial OLS model performs better and which ones it performs worse in. This would indicate that for these boundaries thefts in the shape boundary are influenced by thefts outside of the shape boundary. The boundaries where the Spatial OLS model performed better include Police Districts, Police Beats, Neighborhoods, and State Senate districts. The shapes where it performed worse include Congressional Districts, Wards, and Ward Precincts (these are mainly political districts rather than geographic boundaries).

As the area of the shapes increase so does the MSE as can be seen in the graph below.




so a more helpful measure will be to normalize the MSE by the average area of the shape. Here the area of the shape is measured in squared degrees based on the latitude and longitude points

that make up the boundary.




This graph shows the best MSE per area is from the Police Beat boundaries performing about 17% better than the next closest boundary. This is an interesting and significant result that could mean a few things. Since these are the boundaries that police use in their patrols it could mean that some police are more successful at deterring crimes than others. For example, imagine that a different police officer was assigned to a specific beat and that the performance of the police officer had a measureable impact in the occurrences of crimes in the future – if the police officer was very good at their job the incidents of theft would decrease in the future whereas is they were poor crime would increase. In this case, we would expect Police Beats to be a very "strong" boundary for incidents of crime within the boundary.

Other boundaries such as Congressional Districts are weak borders of crime incidence. Congressional districts were the largest area that I tested and are more political boundaries that aren't a good indicator of crimes. Ward precincts, the smallest boundary I tested was too small

to have good predictive power. In this case, I believe that the number of boundaries was too large and thus was too sparse to study this dataset.

Another interesting finding was that by adding the Twitter data the MSE of my predictions decreased by 5% on average across the boundaries tested. This was surprising to me and would be an interesting dataset to explore further. Twitter data has the advantage of having a lot of observations and I could potentially enrich it with other measurements common in textual analysis such as the predicted education level of the user. Also, with data collected over a longer timeframe it would be interesting to see how sentiment about certain areas changes and how that is indicative of crimes.

**Conclusion**

In this study, I described a novel way to predict the location of thefts using community input dataset from Chicago's 311 system as well as how people describe places on Twitter. The goal of the paper was to show how a model like this could be effective in predicting the location of crimes and how it could be deployed as a real-time system in a city.

While this model proved effective in Chicago there are some limitations that might not make it applicable to other cities as mix of crimes and how cities report them are different. While in For example while in Chicago 37% of all crimes are theft-related, in Los Angeles the percentage is closer to 60%. Additionally, not all cities have a 311 system or make that data publically available and this is a key input to my prediction model.

I believe that further research could be done in this area to improve this model by adding more variables like weather or seasonally adjusting the data to include more observations in the training set. Also, while my OLS model provides a clear baseline for analysis, further experimentation can be done by combining variables or deriving new measure from the datasets such as how certain variables have changed over time, however I believe that this is a promising start and demonstrates an area of research where more academic research can have a positive impact in society.

**References**


Quetelet, Adolphe. *Sur l'homme et le développement de ses facultés ou essai de physique sociale*. Vol. 1. Bachelier, 1835.

Becker, Gary S. "Crime and punishment: An economic approach." *The Economic Dimensions of Crime*. Palgrave Macmillan UK, 1968. 13-68.

Ehrlich, Isaac, and George D. Brower. "On the issue of causality in the economic model of crime and law enforcement: Some theoretical considerations and experimental evidence." *The American Economic Review*77.2 (1987): 99-106.

Glaeser, Edward L., Bruce Sacerdote, and Jose A. Scheinkman. "Crime and social interactions." *The Quarterly Journal of Economics* 111.2 (1996): 507-548.

Gottfredson, Michael R., and Travis Hirschi. *A general theory of crime*. Stanford University Press, 1990.

Glaeser, Edward L., et al. "Big data and big cities: The promises and limitations of improved measures of urban life." *Economic Inquiry* (2016).

Kang, Jun Seok, et al. "Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews." *EMNLP*. 2013.

Courneya, Kerry S., et al. "*Randomized controlled trial of exercise training in postmenopausal breast cancer survivors: cardiopulmonary and quality of life outcomes." Journal of clinical oncology 21.9 (2003): 1660-1668.*

*Mohler, George. "Marked point process hotspot maps for homicide and gun crime prediction in Chicago." International Journal of Forecasting 30.3 (2014): 491-497.*