The Scene of the Crime

Predicting Theft Locations in Chicago

Benjamin Rothschild
MACS 30200
bnroths@uchicago.edu

RESEARCH QUESTION

How accurately can a model predict the locations of thefts, burglaries, and robberies in Chicago?

INTRO

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 has shown that mathematical models can better predict crime than dedicated crime analysts and by using these models for policing and preventative measures, Los Angeles was able to decrease theft-related crime by 7.4% in a controlled experiment (Courneya 2003).

In this study I perform a similar analysis in Chicago while using novel datasources 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 police departments and cities priorizitze their resources.

DATA

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 - Gangs	2,243	4	
311 Alley Lights Out - No Gangs	4,333	8	
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 Public Helath Food Protection Program
Food Inspection - Pass w/Condition	5,674	11	
Food Inspection - Fail	3,342	6	
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.

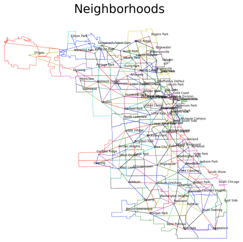
Each data point in the table has a timestamp and latitude and longitude attached.

A point-in-polygon calculation is made to place an event into in a shape boundary in the maps below to be used in the regression model.

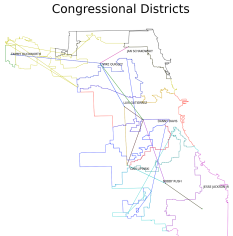
Some variables such as liquor liscenses are included without a time variable in the regression model.

Below are maps of all tested boundraies with a line connecting their queen-neighbors. A more detailed explination of the boundaries and methedology can be found in my paper.

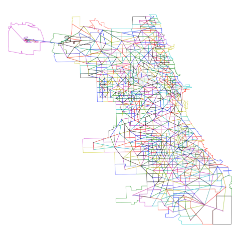
Neighborhoods



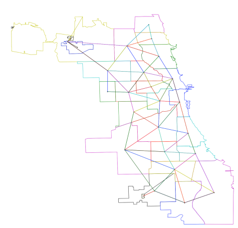
Congressional Districts



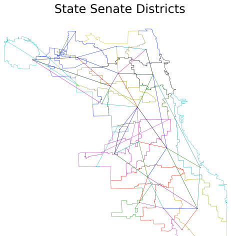
Police Beats



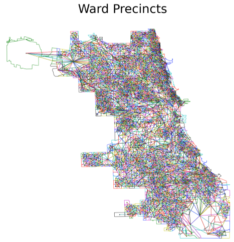
Police Districts



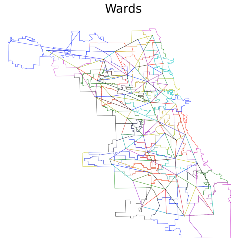
State Senate Districts



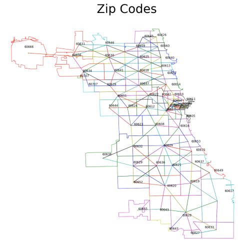
Ward Precincts



Wards



Zip Codes



METHODS

OLS Regression

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

Spatial OLS Regression

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

j	0	i
Knowledge Start	Knowledge Cut-Off	Prediction

I perform two regressions, an OLS regression which I use as a baseline and a Spatial OLS regression which includes data from spatial neighbors.

In the formula's at left (t,j) denotes the count of an events type in shape j at time t.

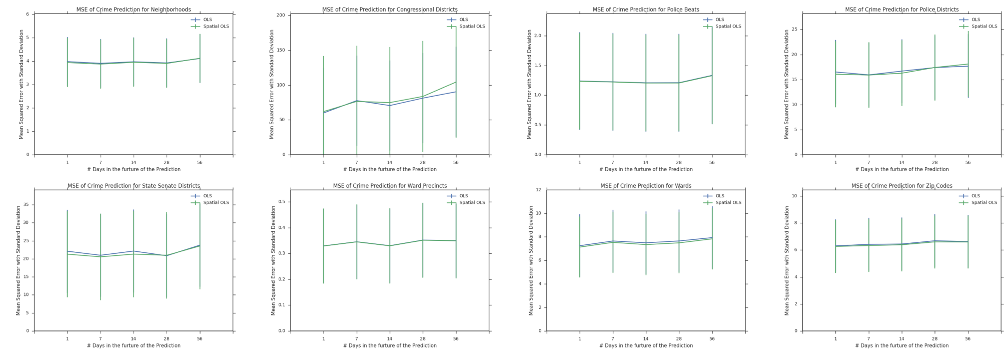
To determine the spatial weights I used queen-contiguity weighting matrix.

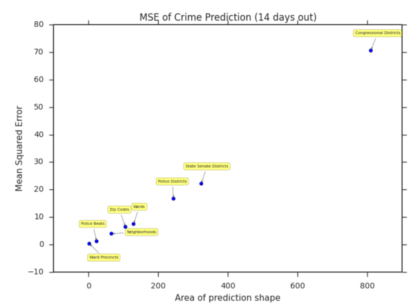
For my predictions, I used the following cross-validation algorithm to calculaute the Mean Squared Error:

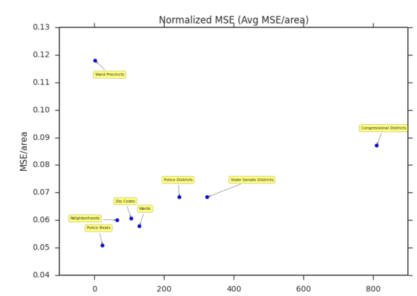
1. choose 200 random days within the time-period
2. train model on previous [j, 0] days
3. make prediction for day i

*I tried several values for j and found that in most cases a 28-day "knowledge window" gave a minmum or sufficiently minimum MSE. This methedology is explained more in my paper.

RESULTS & CONCLUSION







The graphs above compare the MSE of the OLS Model vs the Spatial OLS model. In general the results of the Spatial OLS Model wer not very different than the standard OLS Model, in some cases performaing better (Police Districts) and others performing worse (Congressional Districts). This shows that crime within some boundaries are more influenced by their neighbors (Police Districts) while for other boundaries their is less influence (Congressional Districts)

Comparing the MSE to the area of the shape used shows the tradeoff between area and MSE when making a prediction.

Normalizing the MSE by the area of the predicting region can show us what "strong" boundaries for crime prediction. This third graph shows that Police Beats, Wards, Neighborhoods and Zip Codes are the strongest boundaries for crime while Ward Precincts and Congressional Districts are weaker boundaries for crime.