Benjamin Rothschild
MACSS 30200
Data, Methods, Initial Results

Research Question

How accurately can I predict the location of a theft in the City of Chicago?

Data

I have compiled several data sources for use in my model. Most of the data comes 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 also 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.[1] 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 Data Portal. The specific datasets I am using are:

- Chicago Crime Dataset
- 311 Service Requests
- Red Light Tickets
- Food Inspections
- Liquor Licenses
- Building Violations

A description of each dataset is listed below. 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.

Chicago Crime Dataset 2001 – present

This dataset reflects reported incidents of crime (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 will further divide these crimes into 5 types described below.[2]

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

[2] 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

Primary Type	Count	Examples
Theft	125,034	Burglary, robbery, motor vehicle theft
Violent Crime	97,621	Battery, assault, sexual assault, homicide, kidnapping
Non-Violent Crime	47,195	Deceptive practice, gambling, intimidation, obscenity
Property Damage	40,325	Criminal damage, arson
Minor Crime	2,081	Public peace violation
Other	21,886	
Total	334,142	

Data Collected By: Chicago Police Department

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.

Event	Count	Date Range	Description
Graffiti Removal	26,291	1/1/2011- 5/15/2017	Requests to remove graffiti with city “blast” trucks
Vacant and Abandoned Buildings or Cars	1,461	1/1/2015-5/15/2008	Requests to inspect a vacant building
Alley/Street Lights Out	10,561	1/1/2011-5/15/2017	One or more street lights out
Sanitation Code Complaints	12,925	1/1/2011-5/16/2017	Complaints such as overflowing dumpsters or garbage in the Alley

Data Collected by: City of Chicago

Red Light Tickets

Red light tickets are compiled by the Chicago Tribune and published monthly. These only include tickets that are given through the Red-Light Camera system where cameras are installed at intersections and given to people automatically.

Event	Count	Date Range	Description
Red light tickets	86,817	1/1/2007- 5/15/2017	Red light tickets given by cameras

Data Collected by: Chicago Tribune

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:

- **Pass**: 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.

Event	Count	Date Range	Description
Food Inspections	146,955	1/1/2010-5/15/2017	Food inspections and results

Data Collected by: Chicago Department of Public Health's Food Protection Program

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.

Event	Count	Date Range	Description
Current Liquor Licenses	905	Present/active	Current Active liquor licenses

Data Collected by: Department of Business Affairs and Consumer Protection, City of Chicago

Building Violations

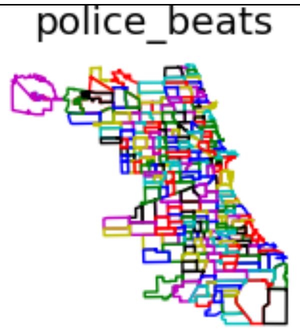
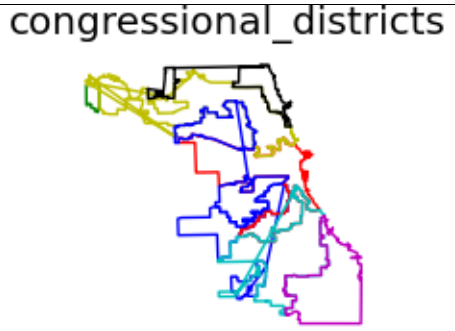
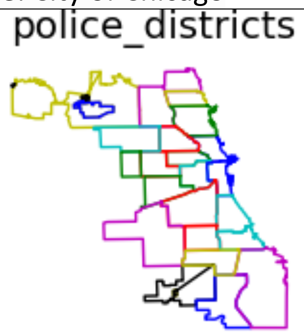
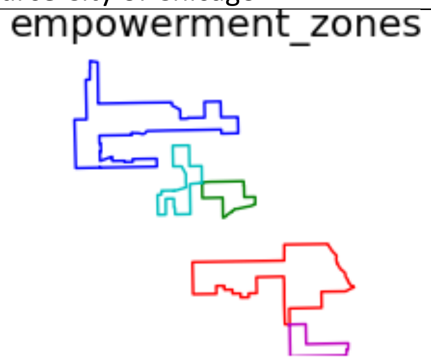
Violations are issued by the department of Buildings and can be for items such as failed maintenance or if a specific building is unsafe or not up to code.

Event	Count	Date Range	Description
Building Violations	27,032	1/1/2006-5/16/2017	Building code violations

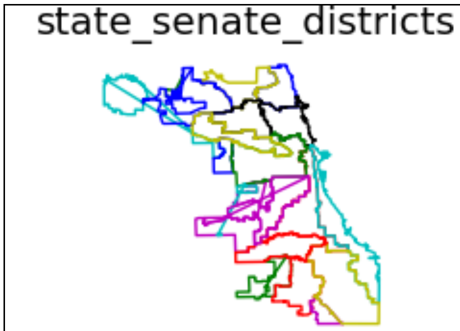
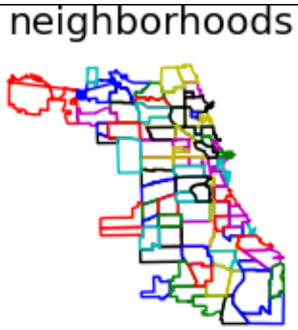
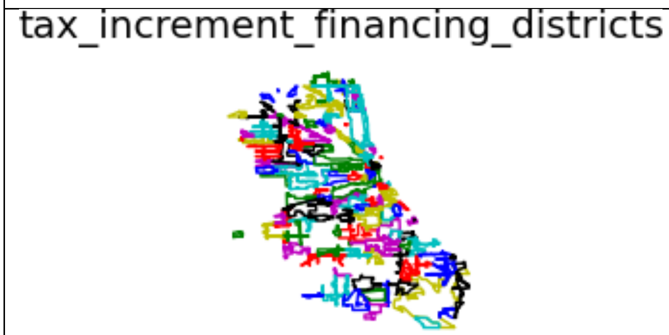
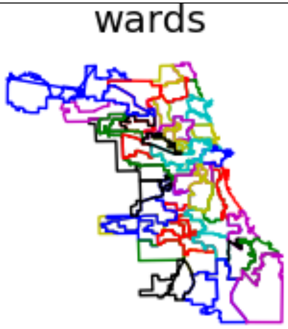
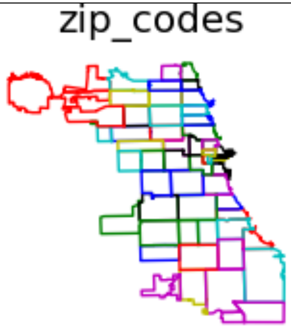
Data Collected by: Department of Buildings

Region Boundaries

Since I am doing a spatial analysis, I wanted to test multiple region boundaries in my spatial regression model to see what boundaries gave the best prediction. I identified 10 different possible boundaries which vary in their size and boundary type. A map and description of each boundary tested is below.

 Current police beat boundaries in Chicago. Source: City of Chicago	 U.S. Congressional district boundaries in Chicago. Source City of Chicago
 Police district boundaries in Chicago. Source: City of Chicago	 Empowerment Zone boundaries in Chicago. The Empowerment Zones/Enterprise Communities program (EZ/EC) is a Federal, State, local government partnership for stimulating comprehensive renewal--particularly economic growth and social development--in distressed urban neighborhoods and rural areas across the nation.[3] Source: Department of Housing and Urban Development

[3] More information about this program can be found at http://www.hud.gov/offices/cpd/economicdevelopment/programs/rc/

 State of Illinois Senate Congressional district boundaries in Chicago. Source: City of Chicago	 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
 Tax Increment Financing (TIF) district boundaries in Chicago. Source: City of Chicago	 ward_precincts Ward precincts, which function as local election districts, in Chicago. Source: City of Chicago
 Ward boundaries in Chicago from May 2015 onward, corresponding to the dates when a new	 ZIP Code boundaries in Chicago. Source: City of Chicago

City Council is sworn in, based on the immediately preceding elections. Source: City of Chicago	

Methodology

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 hedonic regression is common in regressions involving the crime rate. The simple OLS regression is below where X_t is an array that includes the count of observations of an event during the period. The events I include in X_t are:

- crime_theft
- crime_violent
- crime_non_violent
- crime_property_damage
- crime_minor
- crime_other
- building_graffiti
- building_vacant
- building_lights
- building_sanitation
- red_light_tickets
- food_fail
- food_pass
- liquor_liscenses

$$\text{Crime}_t = \alpha + \beta X_t + \epsilon_t$$

Next, I will add spatial weights to my model by spatially lagging my exogenous regressors. Now, each observation is the count of events (listed above) that happened during time period *t* in region *i*

$$\text{Crime}_{it} = \alpha + \beta X_{it} + \delta \sum w_{jit} X_{it}' + \epsilon_{it}$$

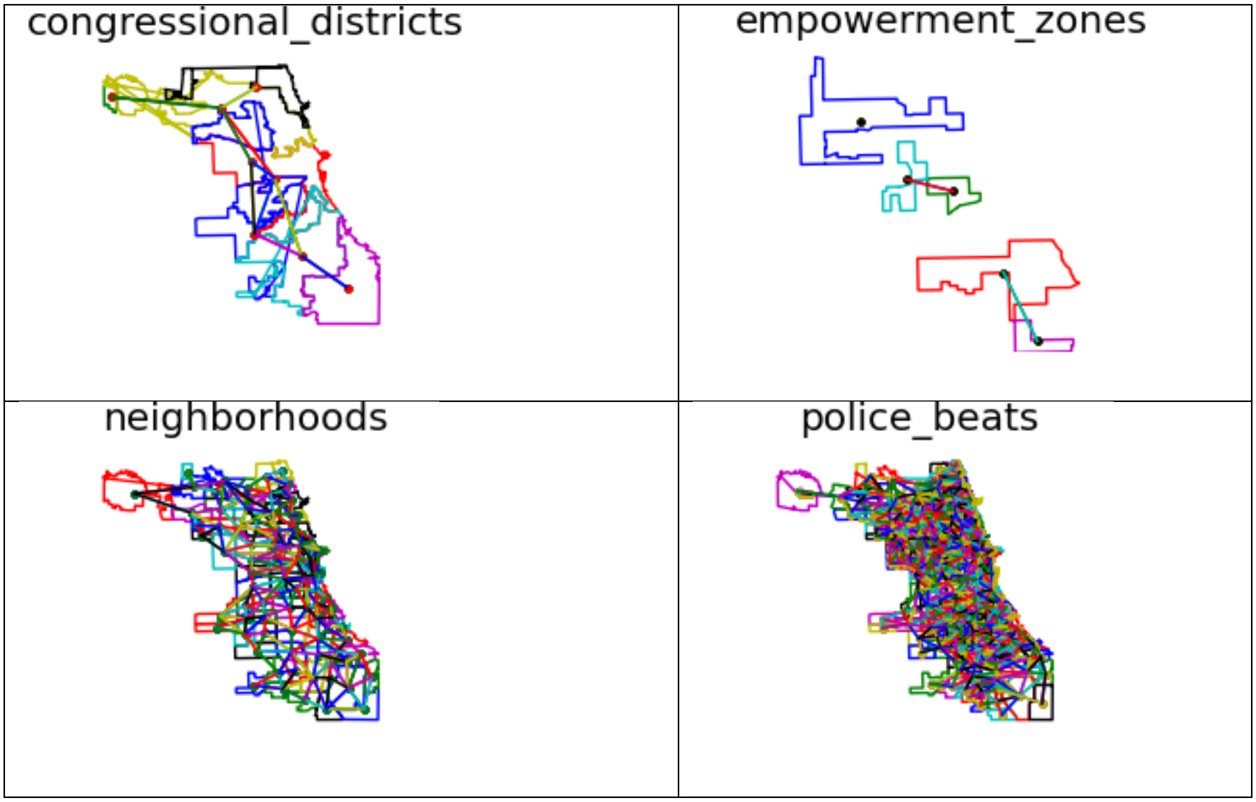
In this model w_{ijt} is a spatial weight matrix. Spatial weights are mathematical structures used to represent spatial relationships. Given two locations *i* and *j*, a spatial weight w_{ij} is a defined geographic relationship between the locations. In many cases, it is some measure of proximity. I decided to use the *Queen Weight* which reflected adjacencies with a binary indicator. Other possible weighting schemes include, rook weights, bishop weights, distances, k-nearest neighbor weights. Essentially what the spatial weight does is average the observations for the surrounding regions and includes the result as a variable in the regression. In the maps below I present the previous boundaries I had with a line connecting each shape to its neighbor.

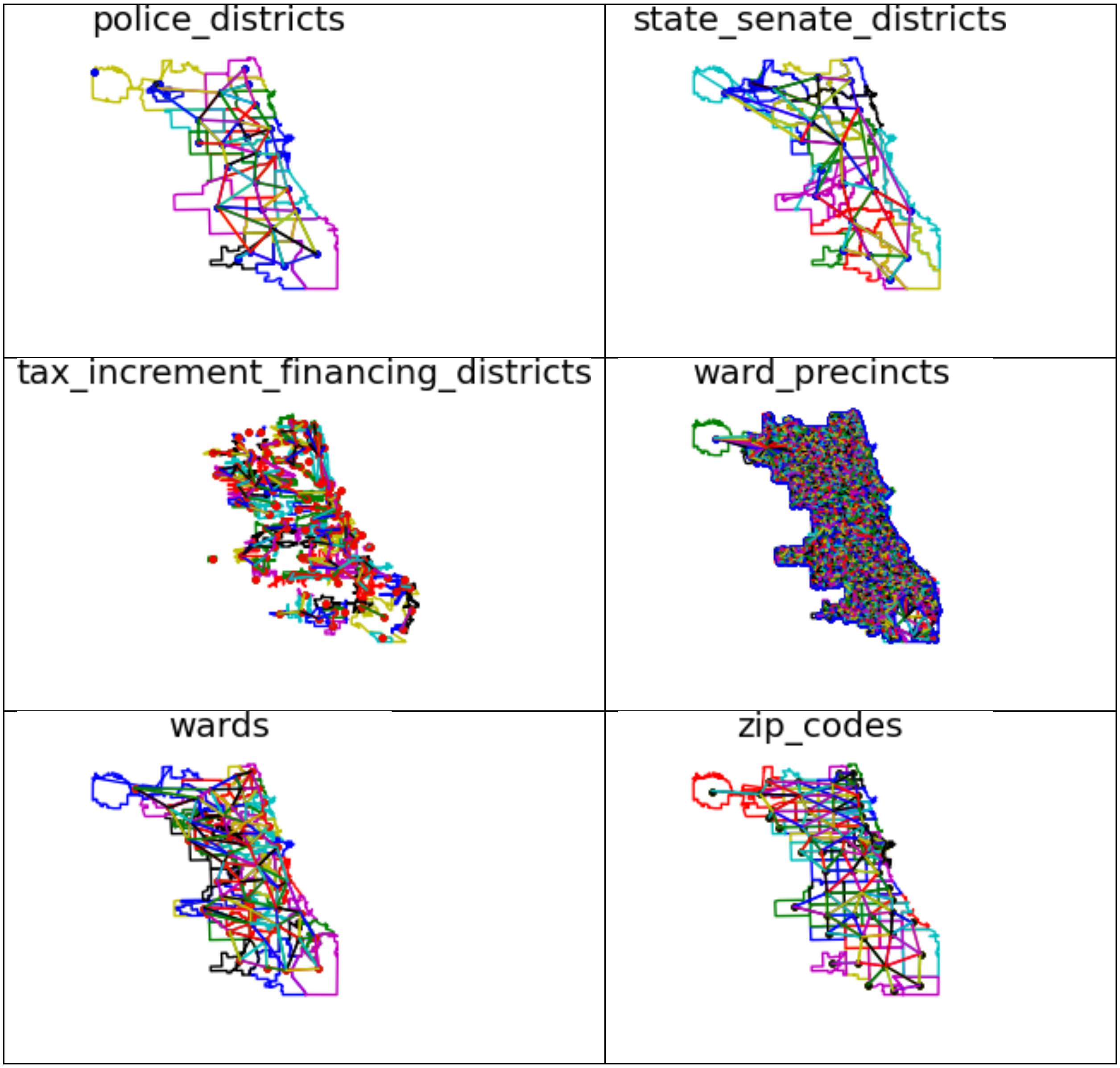
Lastly, I will try a model that will also spatially lag my endogenous variable of neighboring regions. This model looks like this

$$Crime_{it} = \alpha + \lambda_j \sum w_{ij}(Crime_{it} + \beta X_{it}) + \epsilon_{it}$$

Initial Results

So far, I have started with my simple OLS model and created my spatial weighting matrices from the boundary files described above. A visualization of my spatial weights is below. Each boundary is the same as before and I added a line connecting a boundary to its neighbors included in the spatial weight.





police_districts

state_senate_districts

tax_increment_financing_districts

ward_precincts

wards

zip_codes

I also started to run my spatial regression but it is not good enough to share yet. My next steps are to load the shapefiles into PostGIS and transform my event counts to counts per region. I will then use this dataset in my spatial regression.