Multivariate spatial analysis of crime in the city of Chicago

Location: City of Chicago

Level of detail: Census tracts

Coordinate Reference System (CRS): NAD 1983 State Plane Illinois East (EPS102671)

Introduction

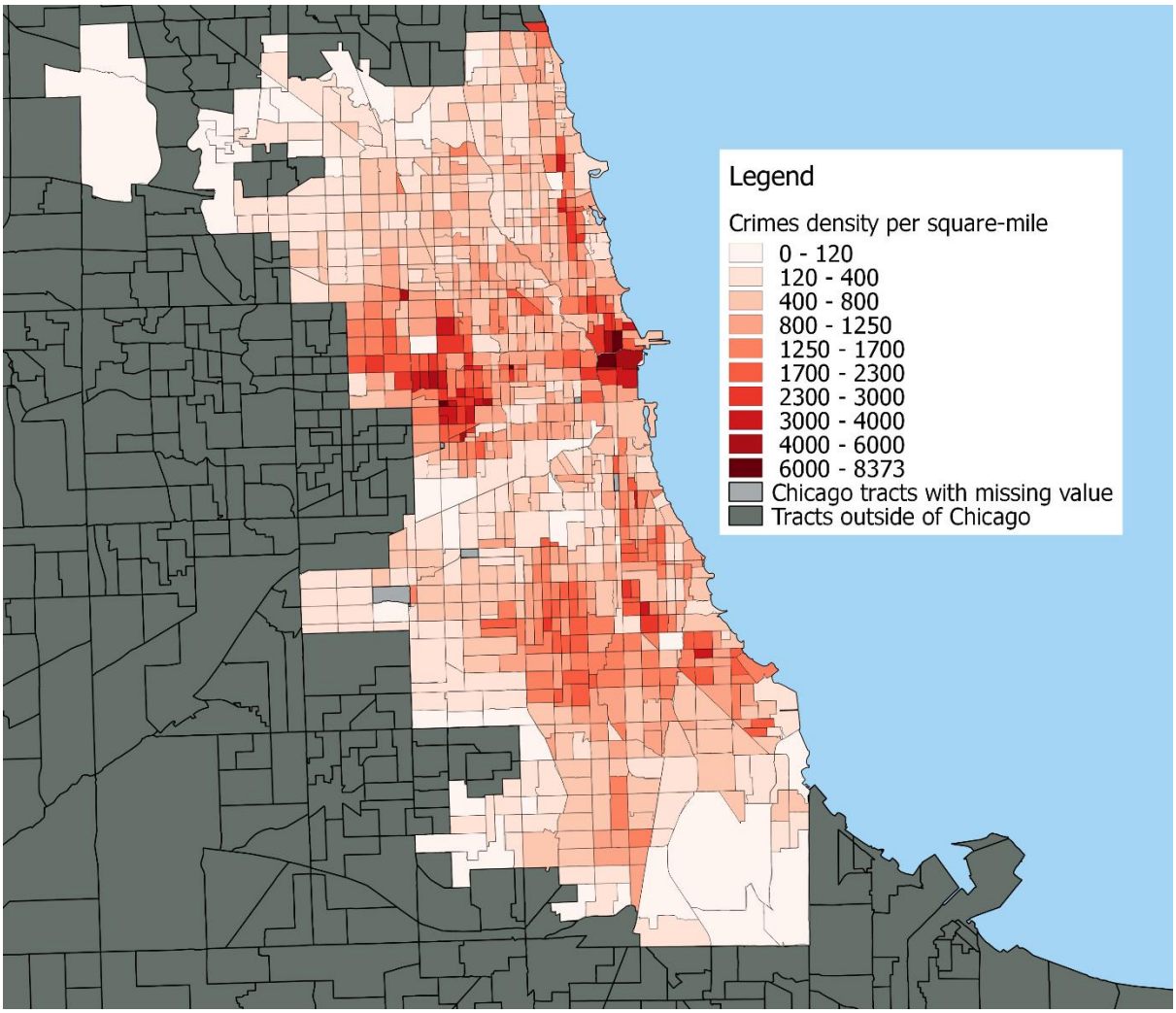
This study aims to explain crime in Chicago census tracts using economic and demographic variables of college education, income, and unemployment rate. We first explore crime by creating a choropleth map of the crime density to identify potential clusters of criminality, and by using centrographic statistics to better understand the spatial dispersion of selected types of crimes across the city. Then we plot choropleth maps of college education, income, and unemployment rate in Chicago to visualize if they are spatially correlated with crime. We hypothesize that these 3 explanatory variables are statistically significant in explaining crime, and more specifically, that unemployment percentage is positively correlated with the number of crime incidents per tract whereas income and college education are negatively correlated with this number. To formally verify this hypothesis, we compare the results of an ordinary least squares regression (OLS) with a spatial regression that accounts for spatial dependence between tracts.

Dataset

I put together the dataset used in this study by joining three files: the first one is a shapefile containing the census tracts of the city of Chicago based on the last census in 2010[1]. The second one is a .csv table which lists all crimes committed in Chicago in

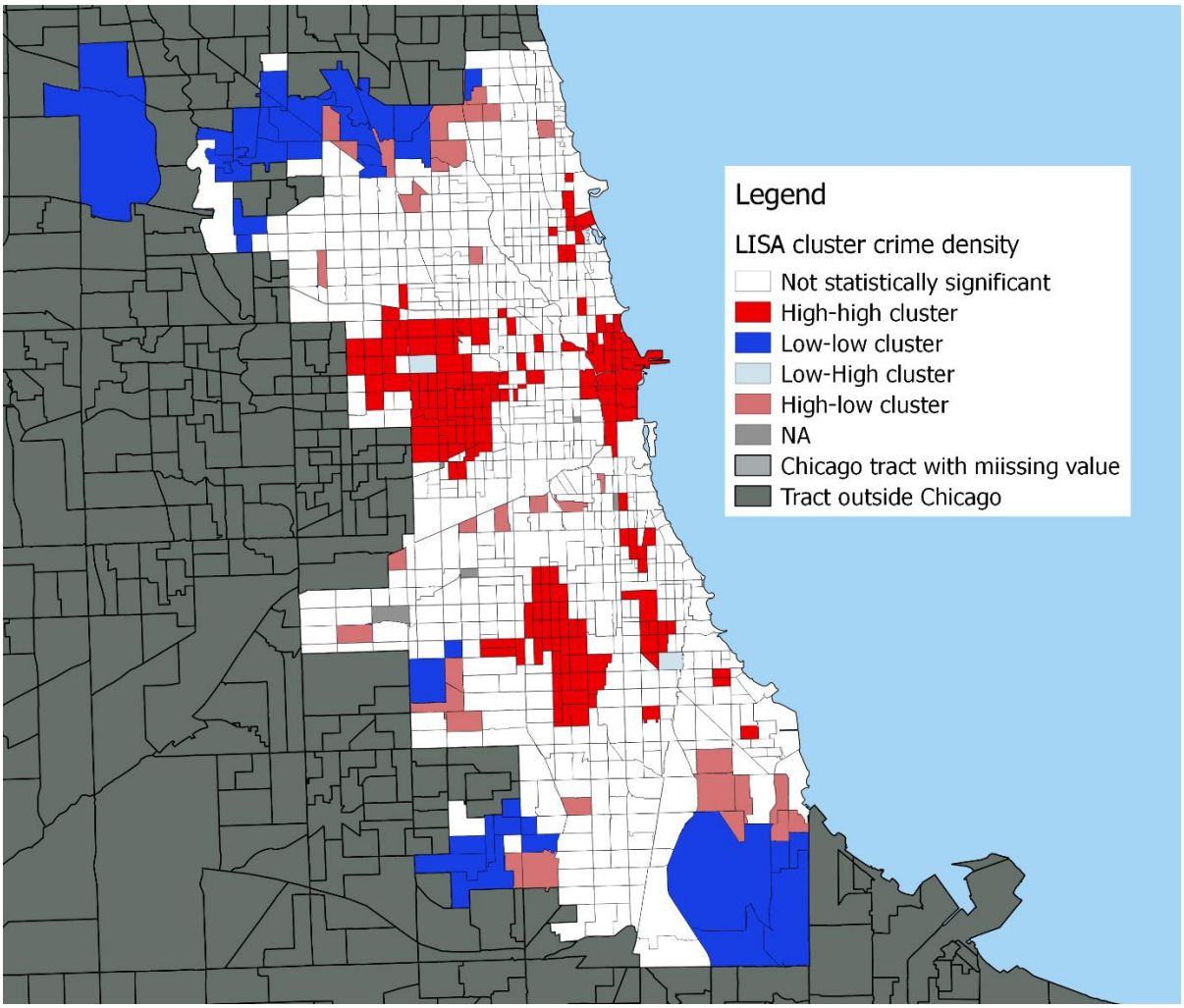
2018 and it is updated live by the Chicago Police Department[2]. The third one is a .csv file from the National Historical Geographic Information System (NHGIS)[3] which includes total workforce and counts of people unemployed in every US census tract as of the latest official survey in 2016. I projected the cloud of crime points on the Chicago tracts shapefile in order to get the number of crimes per census tract in 2018. Then I divided the counts by the surface area of each tract in order to obtain a crime density in “crime / square-mile”. Note that surface areas of census tracts are defined based on the number of inhabitants who reside in the tract, thus my crime density is equivalent to a count of crime per inhabitants. Afterwards, I joined this set with the NHGIS table containing the number of people unemployed per tract and I divided this feature by the total number of people in the workforce per tract to obtain the unemployment rate per tract. Finally, I obtain a dataset that includes all the Chicago census tracts as of 2010, the crime density per tract using 2018 data to date, and unemployment percentage per tract as of 2016.

Choropleth map of crime density



The choropleth map above was plotted using the crime density variable. Natural breaks were selected to segment the range of density because they gave the best results in terms of visually identifying areas of high and low criminality. 10 segments were used (more than the standard 5 segments) because the wide range of data spans from 0 to 8373 crimes per square-mile and requires more segments to achieve appropriate level of granularity. The crime map suggests that there are three crime hotspots (clusters of high criminality) in Chicago: one in downtown by the shore, one in the west side of the city and one in the south side.

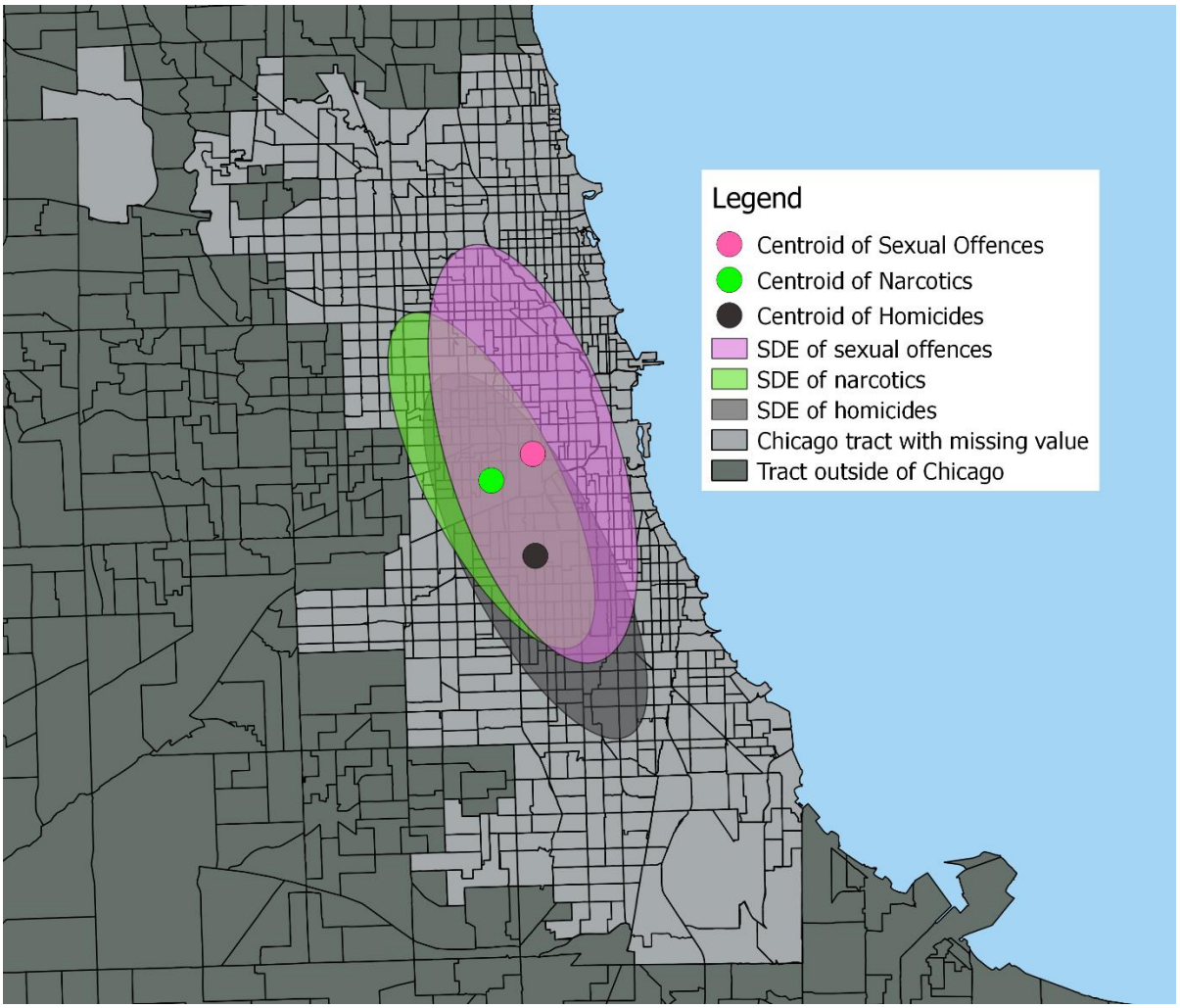
LISA cluster map of crime density



Based on the previous choropleth map of crime we suspect that there are three clusters of crime in Chicago, in West Side, South Side, and Downtown. The LISA map above was plotted using the crime density variable to verify this hypothesis using a queen weight matrix. The matrix was of order two as we believe that the clusters of high criminality have at least two layers of contiguous tracts. The results shown on the map validate our conjecture as we identify three statistically significant clusters of high criminality located around the west side, south side and downtown area of Chicago. The results also show two clusters of low criminality in Chicago O'Hare's international airport and in Wolf Lake memorial park. This is typically expected because airports are

safe areas where many police officers and private companies ensure safety of travelers, and Wolf Lake area has a park and a golf course which are gated areas with low pedestrian traffic.

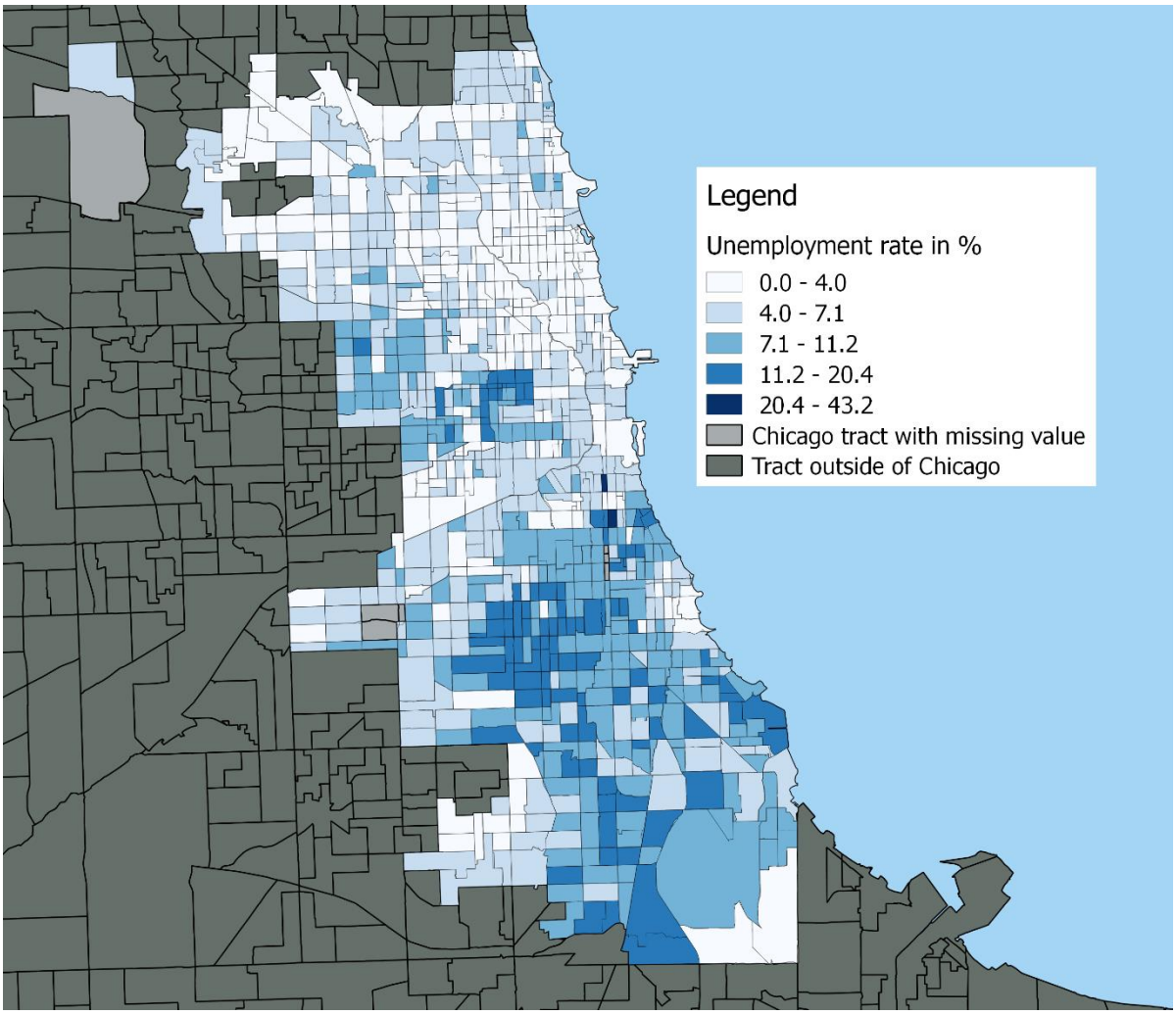
Centroids and standard deviational ellipses for three common types of crimes.



The map above shows the location of the centroids of three of the most frequent types of crimes in Chicago in 2018: homicides, sexual offences and narcotics use. These groups are respectively plotted in black, pink and green. Moreover, we added standard deviational ellipses around each group centroid to gain further insights into their dispersion across the city. We observe that the black centroid of homicides is located near South Side which makes sense as gang violence is highly prevalent in this area.

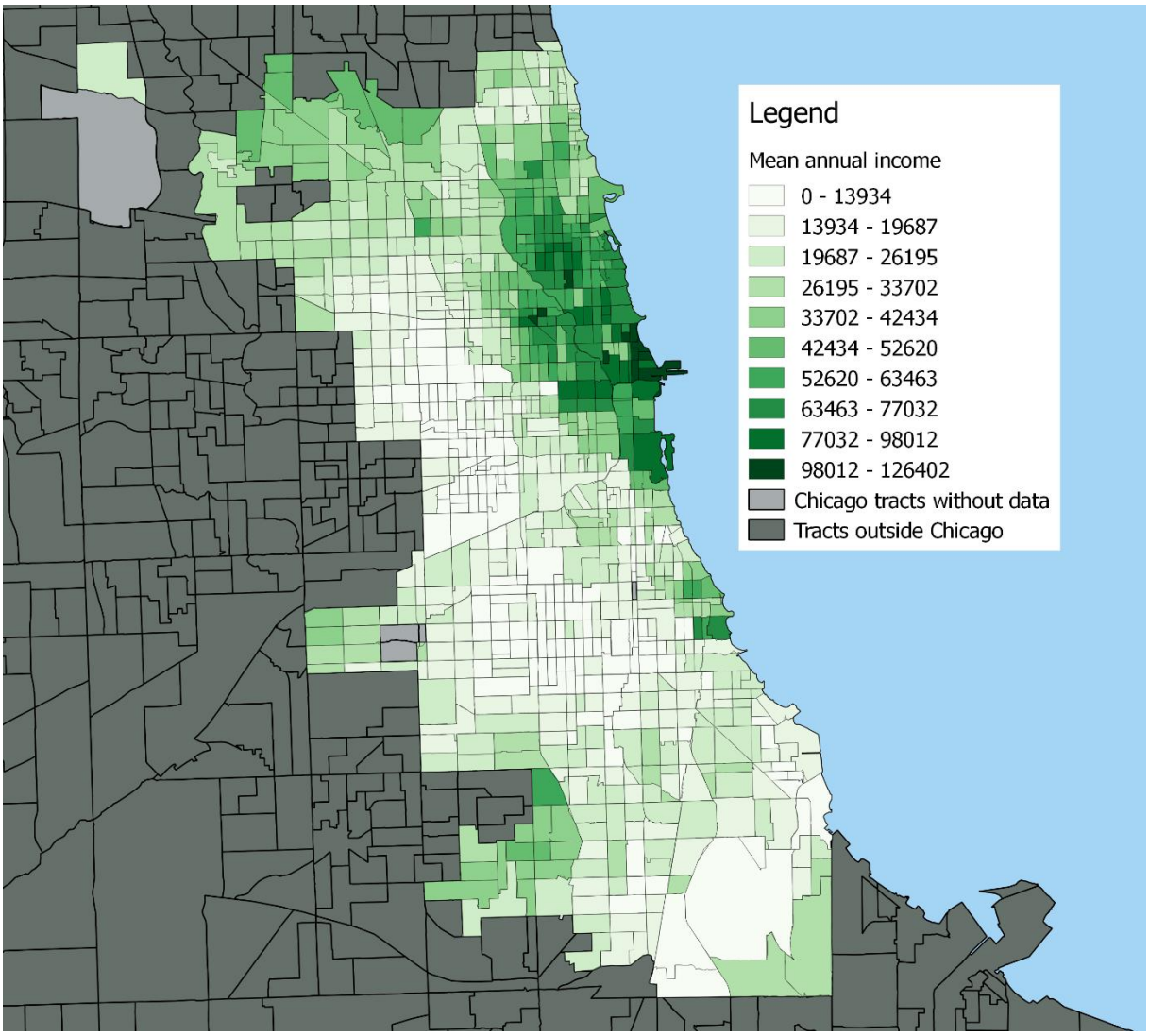
When looking at the pink centroid for sexual offences, we notice a slight shift towards the North East which suggests that these offences tend to occur more frequently near Downtown. We may argue that Downtown nightlife causes more sexual offences, which matches our previous speculation that, in Downtown, wealth and economic dynamic is the root cause of criminality. Finally, we look at the green centroids of narcotics use which is shifted towards the North West near West Side. The shapes and orientations of the three ellipses do not seem to differ from each other and merely seem to follow the outline of the city along the lake.

Choropleth map of unemployment rate



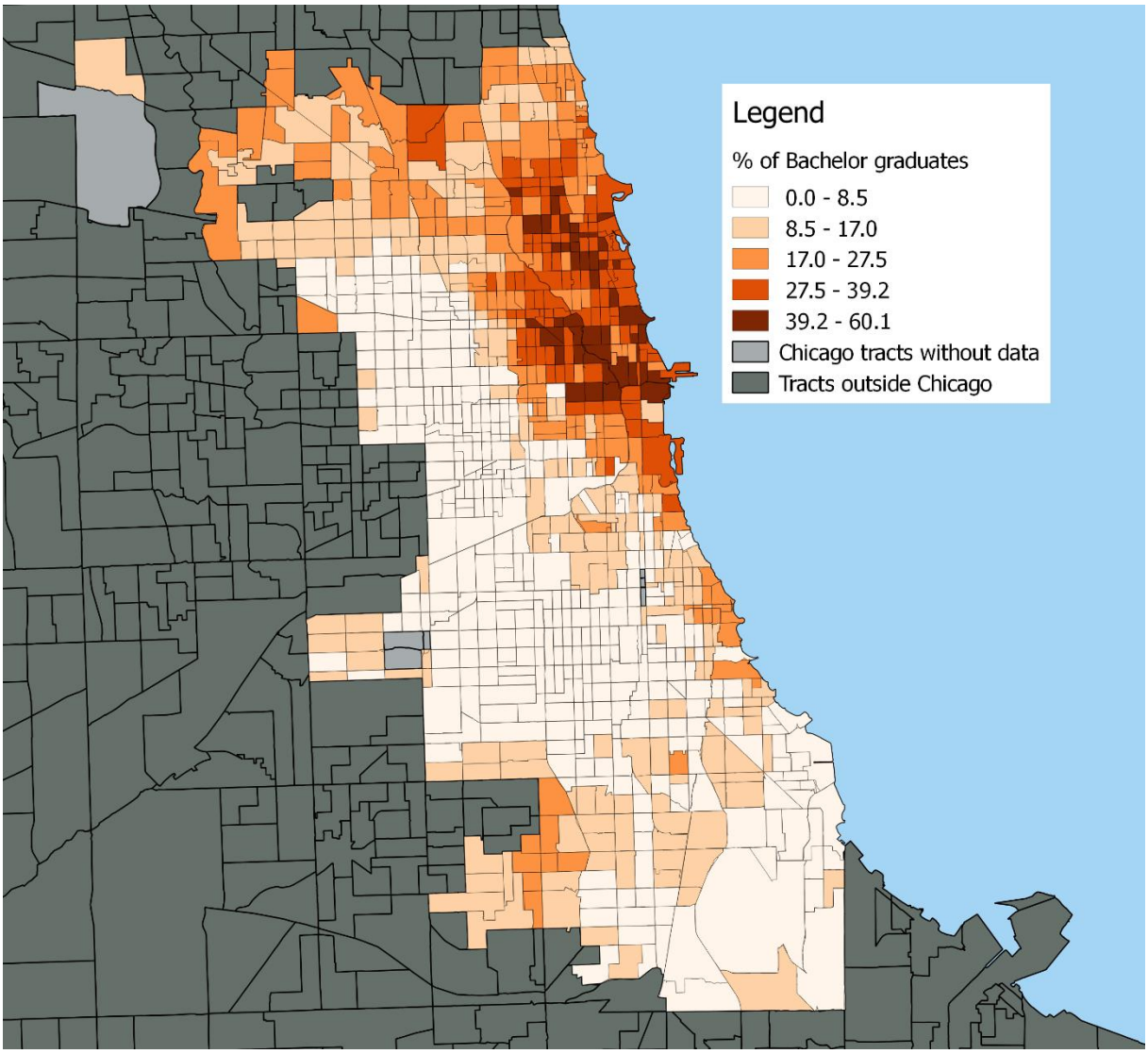
The choropleth map above was plotted using the unemployment rate variable. For consistency, we use natural breaks to plot this data but this time only with 5 segments as the unemployment range is much smaller, from 0% to 43.2%. Unemployment clusters can be seen matching the crime hotspots of West Side and South Side, but the Downtown crime hotspot is invisible on the map. This was expected because Downtown is typically where the highest earners reside, and its cost of living is very high. We may thus draw the conclusion that the types of crime in West Side and South Side must be similar to each other, but both different from the crimes observed downtown. We may further argue that crimes in South Side and West Side are related to the poor economic situation of these neighborhoods while crimes in Downtown are on the contrary triggered by the wealth and dynamics of this area.

Choropleth map of income



The choropleth map above was plotted using the mean annual income variable. We use natural breaks with 10 segments given that the income range is large, from $0 to $126,402. As previously mentioned, there is a cluster of higher earners residing in Downtown while the crime hotspots of West Side and South Side form areas of low income. Therefore, income seems correlated with crime everywhere except in Downtown.

Choropleth map of education



The choropleth map above was plotted using the proportion of residents listing bachelor's degrees as their highest academic achievement. Natural breaks are used to plot this data with 5 segments given the small range of data, from 0% to 60%. The map obtained looks like the two previous ones: college educated people reside Downtown and not in West Side nor South Side. Therefore, education, seems correlated with crime everywhere except Downtown.

Ordinary least squares (OLS)

First, we run an OLS regression in R. The model seeks to explore the dependent variable “Count_crime” which is a count of crime incidents per tract in Chicago from January to September 2018. The independent variables are “Income” (the mean annual income per tract), “Unemployment” (a count of unemployed people per tract) and “Bachelor” (a count of college graduates per tract). We also add a control variable “Land area”, the surface area in m² of each tract, because we expect that larger tracts have more crime incidents than smaller ones. OLS returns the coefficient estimates below.

Name	Estimate	Standard Error	P-value
Intercept	440.4	28.70	0.0000***
Bachelor	-0.2805	0.06036	0.0000***
Income	-0.2456	0.04159	0.0000***
Unemployment	0.1588	0.07853	0.0435*
Land area	0.0000515	0.00001030	0.0000***

We notice that all predictors, namely “Unemployment”, “Income”, “Education” and “Land area”, are statistically significant with p-value < 0.05. The p-value of the F-test is essentially 0 so we conclude that our model fits the data better than a model without these variables. The AIC of the OLS is 11,184 and the adjusted R^2 is about 10%. R^2 is low but this is reasonable given that there are only four predictors. Additionally, we observe that our former hypothesis is confirmed by the signs of the estimated coefficients with positive coefficient estimates for “Land area and “Unemployment”, and negatives coefficients for “Income”, “Education”. We interpret them as follows:

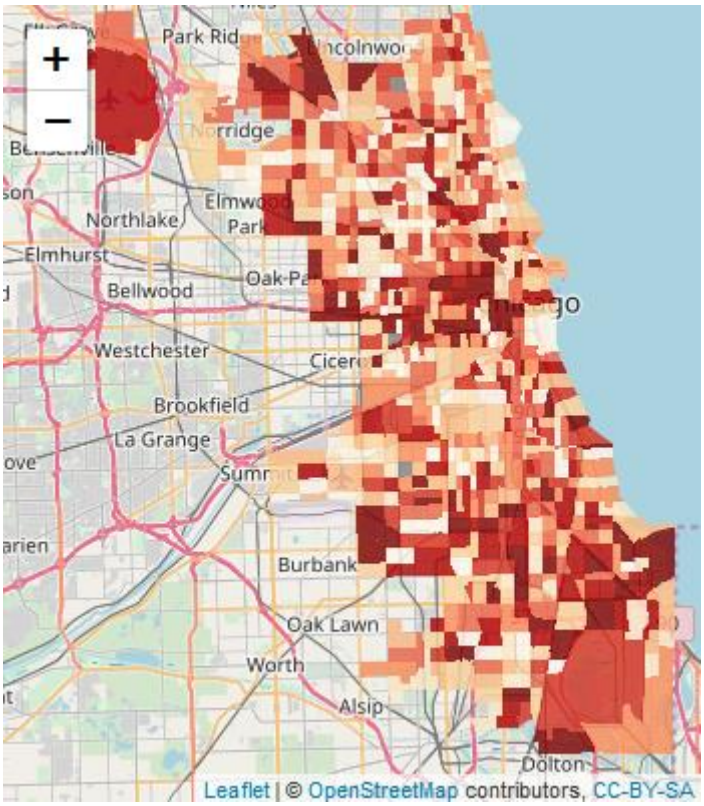
- Increasing by 1m² the land area per tract increases the number of crimes per year in this tract by 0.00005147
- Controlling for land area, the addition of 1 unemployed person in the tract's population increases its annual number of crimes by 0.1588.
- Controlling for land area, an increment of $1 in the tract's mean annual income decreases its annual number of crimes in the tract by 0.2456.
- Controlling for land area, each college graduate in the tract's population decreases its annual number of crimes in the tract by 0.2805.

Spatial autocorrelation model

One of the caveats of the OLS model in the context of spatial regression is that the assumption of residuals independence is often violated. Let's verify if this assumption holds using a double-sided Moran's I test and a visual inspection of the distribution of residuals across the city. To build the Moran's I test we first create a spatial weight matrix of order 1 with Queen contiguity because we expect all connected tracts to have an influence on each other. The test returns the below output.

Observed Moran I	Expectation	Variance	P-value
0.411	-0.00220	0.000396	0.0000***

The null hypothesis of the test is spatial independence and the p-value of the test is essentially zero therefore we conclude that there is spatial dependence in our dataset. This is confirmed by a visual inspection of the residuals which are clustered near South Side and Downtown.



Thus, we need a new model that accounts for this dependency. There are two possible models that could be a good fit, spatial lag or spatial error, so we run Lagrange Multiplier diagnostics to identify which one is the most appropriate. The output of this diagnostics is below.

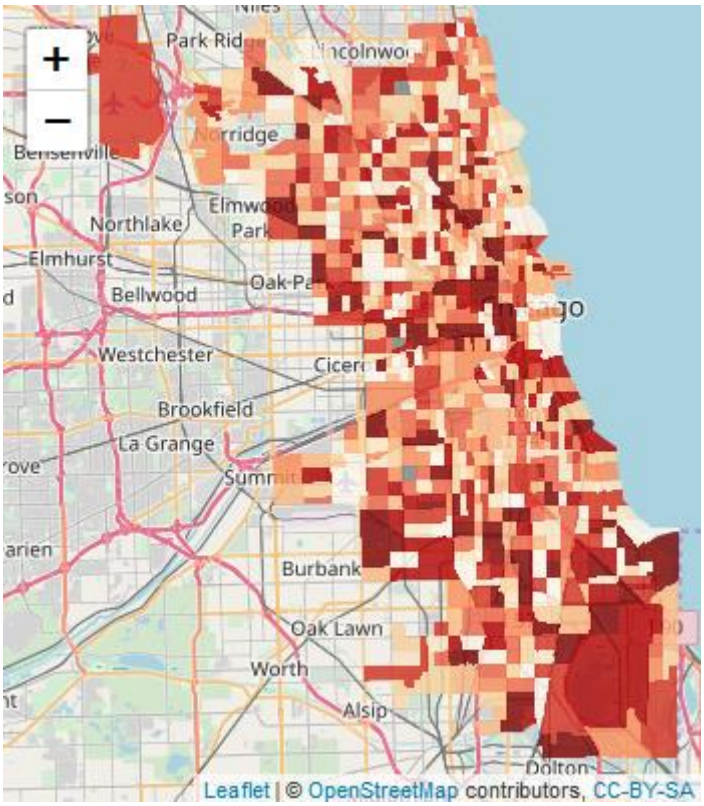
LM Diagnostic	Statistics value	P-value
Error	468.2	0.0000***
Lag	459.92	0.0000***
Robust Error	18.05	0.0000***
Robust Lag	9.77	0.0018***

Both the spatial lag and spatial error tests are significant and thus, inconclusive. We turn our attention to robust forms of the models. We find out that the robust forms are also both statistically significant, but the robust lag model has a higher p-value, so we pick the robust spatial error model. The output of this model is included below.

Name	Estimate	Standard Error	P-value
Intercept	312.73	33.86	0.0000***
Bachelor	-0.129	0.04843	0.007488***
Income	-0.122	0.04694	0.008918***
Unemployment	0.107	0.06361	0.092312*
Land area	0.000072	0.000009322	0.0000***

Lambda, the coefficient of spatially correlated error, is equal to 0.654 and it is statistically significant. This coefficient captures the dependency between tracts, therefore the error model will be improved compared to the OLS model as illustrated by the AIC of 10,902 in lower values than the OLS's AIC of 11,184. The estimates of the predictor's coefficients from the spatial error model are -0.12952 for the education variable, -0.12277 for the income, 0.107 for the unemployment and 0.000072 for land area. The signs of the predictor's coefficients have not changed so our hypothesis still holds with this model. However; all except the control variable land area are smaller in terms of magnitude.

We visualize below the distribution of residuals generated by the spatial error model and notice that the residuals seem less clustered, a sign that spatial dependency has weakened.



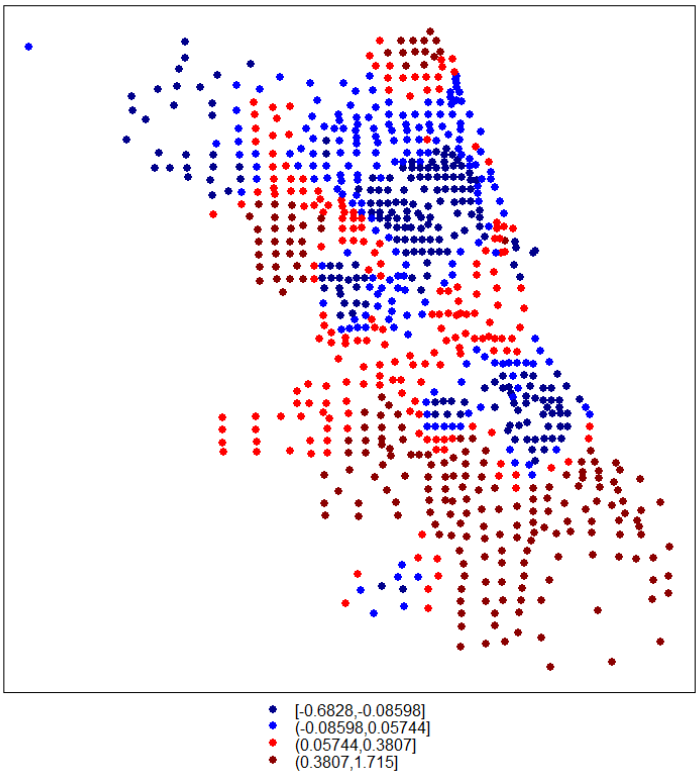
Geographically Weighted Regression (GWR) model

We run a GWR on the Unemployment variable and obtain the below output.

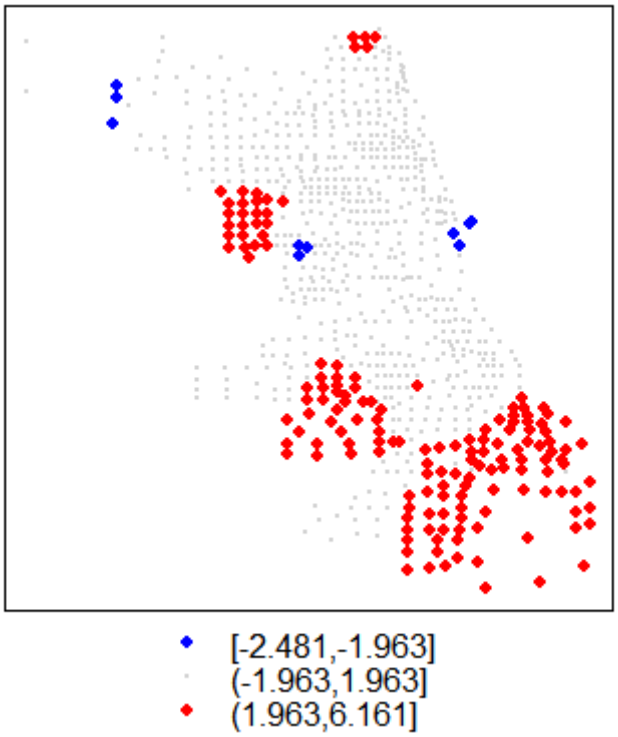
Name	Min	Median	Max
Intercept	-271.16	244.65	726.76
Bachelor	-1.047	-0.05823	0.512
Income	-1.1	-0.201	1.142
Unemployment	-0.682	0.05744	1.714
Land area	-0.00003184	0.0002402	0.00218

The medians of the explanatory variables have the same sign as the OLS coefficient estimates which supports our hypothesis and previous conclusions that unemployment and land area should be positively correlated with crime whereas Bachelor and Income should be negatively correlated with crime. However, in some tracts of Chicago Bachelor and Income appear to have a positive sign while Unemployment and Land

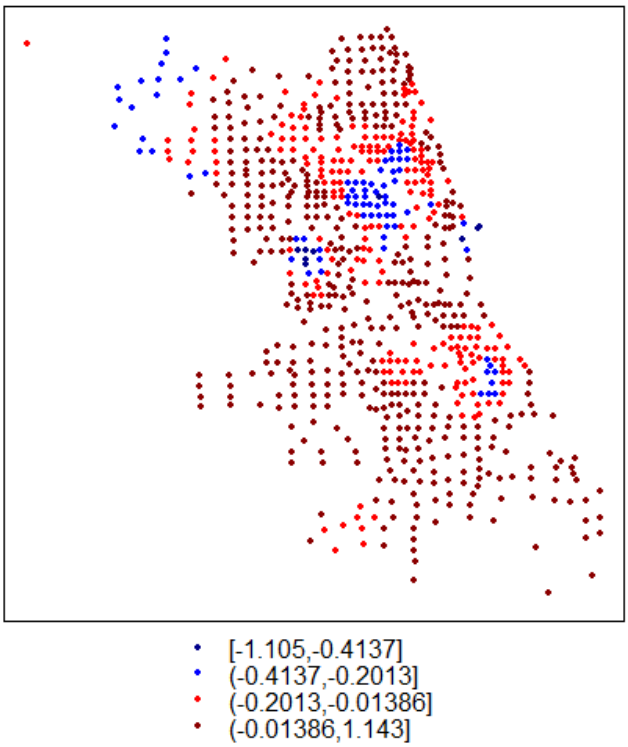
area appear to have negative signs. Therefore, we map the GWR coefficients of each variables to identify these tracts that contradict our hypothesis.



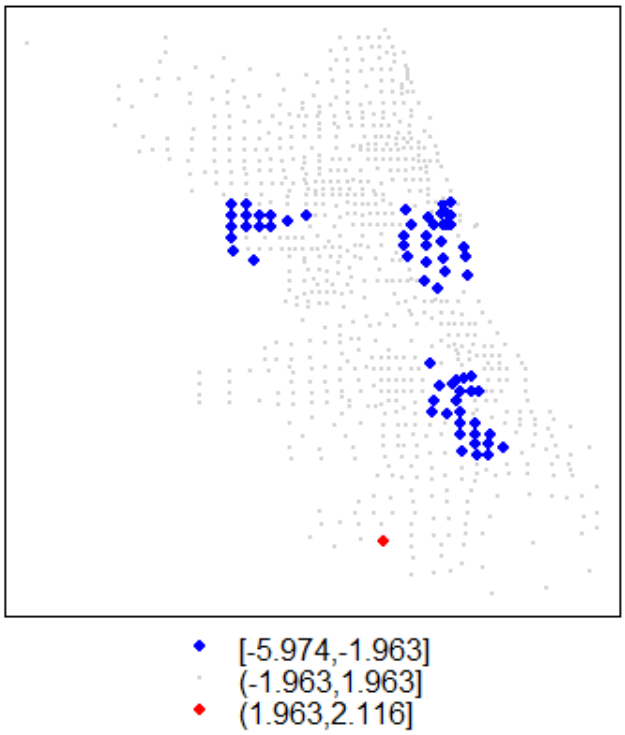
The coefficient for Unemployment plotted above is positive in the South and West side of Chicago and negative Downtown and Uptown. This means that in Downtown and Uptown, more unemployment causes less crimes. However, these areas of the city have the lowest unemployment rate, so we suspect that these local coefficients may not be statistically significant. To verify this, we plot the statistical significance map below.



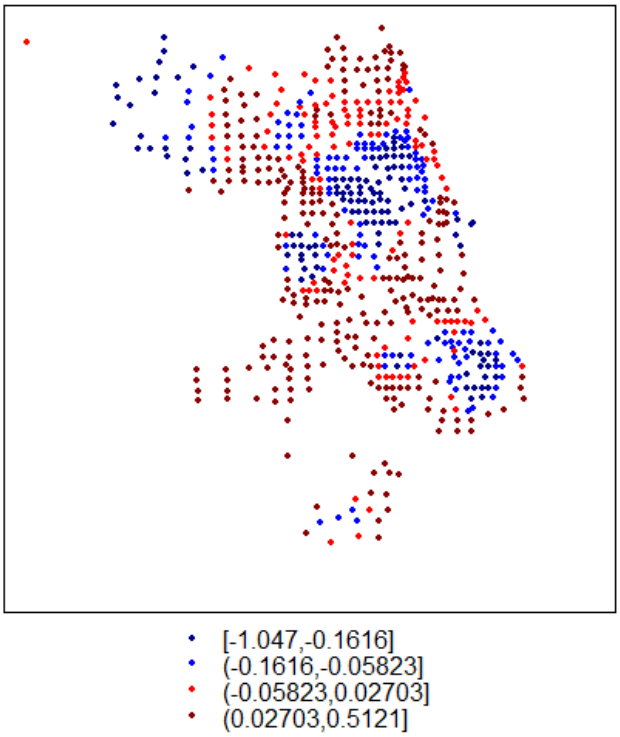
We find that only the West and South sides of the city are statistically significant. It confirms our intuition that Unemployment is only significant in areas where the coefficients are positive, and that negative coefficients are implausible. This result does not contradict any prior conclusions we made but it also does not provide any additional insights because the positive coefficients do not vary a lot among the significant tracts.



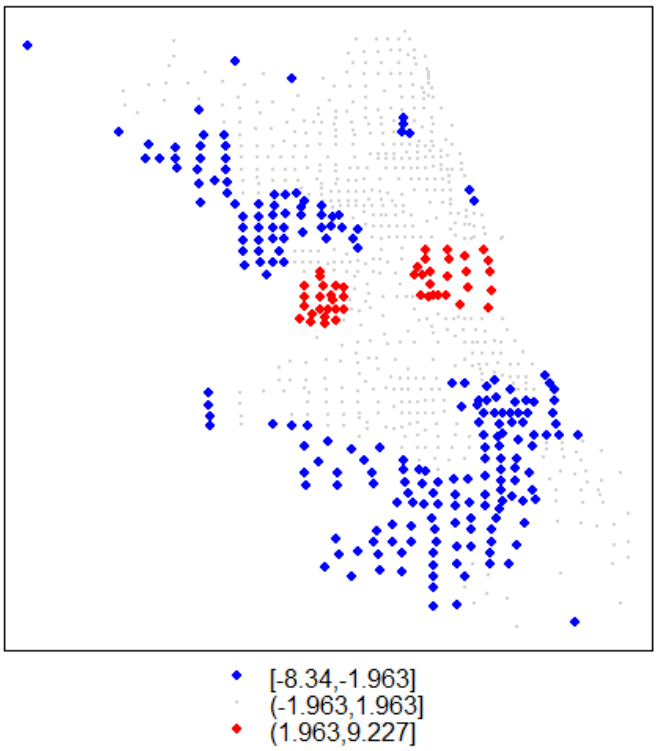
The coefficient for Income plotted above is negative Downtown and in West Side but positive in South Side and around the city limits. This means that in the South, more income causes more crimes. However, the statistical significance map below proves that these areas of the city are not statistically significant, so their coefficient estimates are inconclusive.



The coefficients of “Bachelor” are plotted below. While we expect them to be negative everywhere, we notice that near the South West limit of the city, the coefficients are positive.



Additionally, the significance map further below shows that the coefficients are statistically significant near the South West city limit. Therefore, we conclude that in that area, higher education means more crime. However, there are not many of these tracts and they are right next to another city and they could be seriously influenced by this proximity, so this exception does not contradict our hypothesis.



Conclusion

The OLS, the spatial regression and the GWR confirmed our hypothesis that education, unemployment and income are strong predictors of crime. More specifically, the coefficient estimates of the models showed that unemployment increases the occurrence of crime, while a higher level of education and higher income decrease it. A comparison of the residuals maps proved that the spatial error model is an improvement over the OLS model because it fixes the problem of spatial dependence. Comparing the AICs of the 3 models revealed that the GWR yields the lowest AIC and thus it has the best explanatory power.

Resources

[1] : Chicago census tracts 2010

https://www.cityofchicago.org/city/en/depts/doit/dataset/boundaries_-_censustracts.html

[2] : Chicago crime data 2018

https://data.cityofchicago.org/Public-Safety/Crimes-2018/3i3m-jwuy

[3] : 2016 unemployment data per census tracts

https://www.nhgis.org/